AI ALLIANCE NETWORK

# AI Horizons

## What will AI technologies look like in 10 years?

A Research Project

November 2025

# CONTENTS


About the project 3

Methodology 6

Focus Area 1
Architectures, machine learning algorithms, optimization and mathematics 13

Focus Area 2
Computation for AI 19

Focus Area 3
Data for AI 27

Focus Area 4
Foundation generative models 35

Focus Area 5
Safety, trust and explainability 41

Focus Area 6
Narrow AI 49

Focus Area 7
Control, decision-making, and agentic/multi-agent systems 57

Focus Area 8
Elements of AGI
65

Focus Area 9
Human-machine interaction
73

Focus Area 10
Society in the AI era
81

Conclusion
90

Appendix
91

Authors of the final report
101

Project team
106

# ABOUT THE PROJECT

Today's artificial intelligence (AI) field is shaped by a fundamental contradiction. On the one hand, AI has transformed from a highly-specialized technology into a systemic factor shaping the development of economy, public administration and foundations for sovereignty. On the other, its rapid evolution supported by generative models, multimodality and agentic systems results in a set of very complex problems: from unprecedented requirements to computational resources and power consumption to critical matters of security. It is this duality — the enormous potential fraught with severe limitations — that makes AI one of the top-priority topics. The present day is pivotal moment: today's decisions on technology scaling make a development pathway for many years ahead.

This report was aimed at the comprehensive analysis resulting in the clear delineation of consistent trends and short-term fluctuations.

The study is focused on identification of opportunities and development of reasonable measures for risk mitigation.

The interdisciplinary and international nature of our analysis allowed us to create realistic roadmaps for scientific research, product development and public policy.

10 thematic focus areas of the final report form a consistent and interrelated path for AI development. The study is based on the principle of end-to-end analysis: from source data and existing algorithms to infrastructure support, risk assessment, widespread social and economic effects and, finally, prospects for achievement of AGI. Special attention is paid to a tendency towards AI autonomization directly related to the growth in its agentic nature.

Preparation of materials for the report was based on the comprehensive analysis of the following sources:

**21** foresight sessions held in 2025

**32** in-depth interviews with leading AI experts

Participated in the project:

**270+** leading AI scientists took part in preparation of the final report

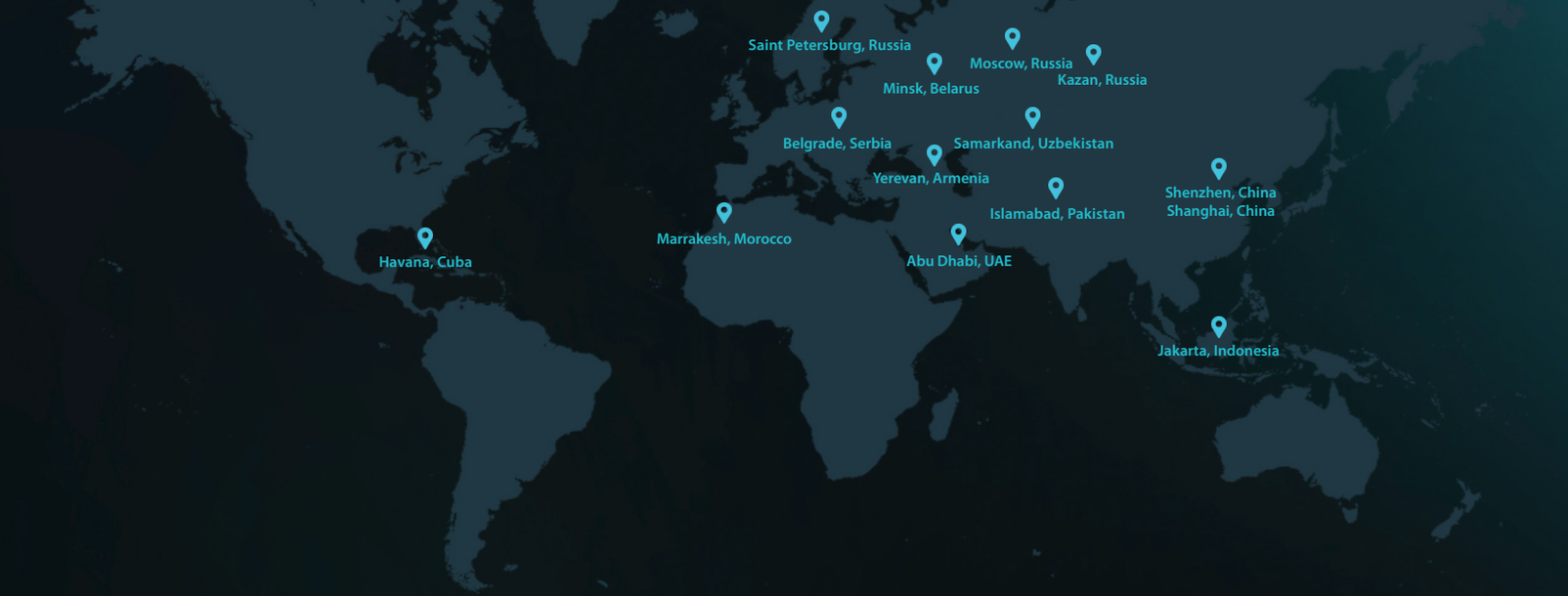
**36** countries — the geography of leading AI scientists

Our research was also based on the comprehensive analysis of public data and industry reports. Such a perspective made it possible to relate the academic agenda with the business practices and specify some consensus points and fundamental contradictions in the AI field, and to identify some bottlenecks in development. The latter include the problems of data quality, insufficiency of computational resources and controllability of complex models. We are convinced that unbiased conclusions may be generated only in the course of broad international professional dialogue, which has become the basis for this report.

# FORESIGHT SESSIONS

The key foresight session was held on 16 June in the Sber Technohub in Saint Petersburg. It gathered together researchers from all over the world to discuss achievements in the AI field and opportunities for international cooperation. The event was participated by 54 leading scientists from 18 countries, including Russia. They worked in 4 groups on 10 topics. Each group was proposed to start from various topics, in order to cover and analyze all proposed fields. Alongside with the work of the groups, there was a series of in-depth interviews, where experts were asked questions about the future of AI and fields of their scientific activity.

Key foresight session of the project
16 June 2025, Saint Petersburg, Russia



# PARTICIPANTS OF FORESIGHT SESSIONS

The scientific foresight sessions brought together over 270 AI researchers from more than 36 countries

# METHODOLOGY

Foresight is a systematic and task-oriented process for construction of knowledge about the future.

As part of the ***science and technology foresight***, the special attention is paid to identification of top-priority areas in the scientific research, potential breakthroughs, as well as assessment of timeframes for solution of particular research problems. This study forms a strong foundation for making of decisions related to planning of research, resources and cooperation between academic institutions, industry and government both domestically and with foreign partners.

The specific nature of the AI field imposes restrictions on foresight studies. AI is characterized by the fast pace of development: development cycles are being reduced considerably, technologies and knowledge about them are emerging and becoming obsolete rapidly, while the profile of scientific research changes considerably every year.

As a result of this, the project was based on expert methods allowing us to get the latest data in a timely manner and quickly adapt to changing realities.

The project was implemented by dozens of high-class experts. They had been formally chosen on the basis of the h-index (not lower than 15), as well as the publication of at least two papers at A-level AI conferences starting from 2020 and/or the considerable number of papers published in top-rated AI journals.

**The foresight study included 3 key stages:**

### 1. In-depth interviews to form the base of independent assessments and proposals

There was a series of in-depth expert interviews held with leading Russian and international scientists aimed at the collection of independent expert assessments by focus areas, as well as exploratory and fundamental research in AI. Structured interviews covered the assessment of current situation, prospects for development and formation of proposals for changes in the list of subareas and research problems formed and updated in the course of the foresight study. This resulted in a dataset of insights, expert assessments and proposals.

### 2. Foresight sessions to form the consensus opinion

One of the key stages of the study was represented by foresight sessions aimed at collection and formation of participatory expert assessments by subareas and research problems in AI. There were 21 foresight sessions, each of which included the work of several working groups formed in accordance with areas of scientific research identified in the course of Foresight 2024. The most important tasks of foresight sessions were the identification of disputable points of view and formation of consensus opinion. The key result of each session was represented by structured conclusions related to updating of the list of advanced studies in AI, specific features of these studies, and assessments of time horizons for achievement of important milestones.

### 3. Final report

The information obtained in the course of expert interviews and foresight sessions was accumulated into a single dataset, after which there was an independent review of conclusions and their adjustment based on additional comments and data.

The validated data provided the basis for the final report. At the level of each of 10 focus areas, the report was prepared by two editors: a leading Russian and a leading foreign scientist in the respective field. Based on the consensus data, the editors form the vision for each focus area of advanced studies with the uniform structure. The report contains summarized results for each focus area, specific research problems, time horizons for their solution and other important insights.

As part of the foresight study, the researchers prepared a set of standard materials (interview guides, standard scenarios for foresight sessions, templates for fill-out forms, presentations for experts, etc.) forming the methodological framework, which makes the future studies much easier.

With a view to prepare for the targeted communication with scientists and clarify the areas of competence, abstracts and full texts of all papers published by Russian and foreign researchers taking part in foresight sessions were analyzed with the use of AI tools.

In order to analyze the discussed topics and scientific challenges, the researchers formed a corpus of audio recordings with the total length of over 60 hours (more than 400,000 words), including materials of foresight sessions and interviews with scientists. This corpus served as the basis for the in-depth AI analysis, which allowed structuring the thematic field and identifying the basic points, which ensured the comprehensive representativeness of opinions given in the course of discussions. Moreover, the foresight sessions were held with the use of a well-prepared system analyzing recommendations of potential academic interactions based on the analysis of publication profiles of various scientists.

# PYRAMID OF FOCUS AREAS IN AI RESEARCH

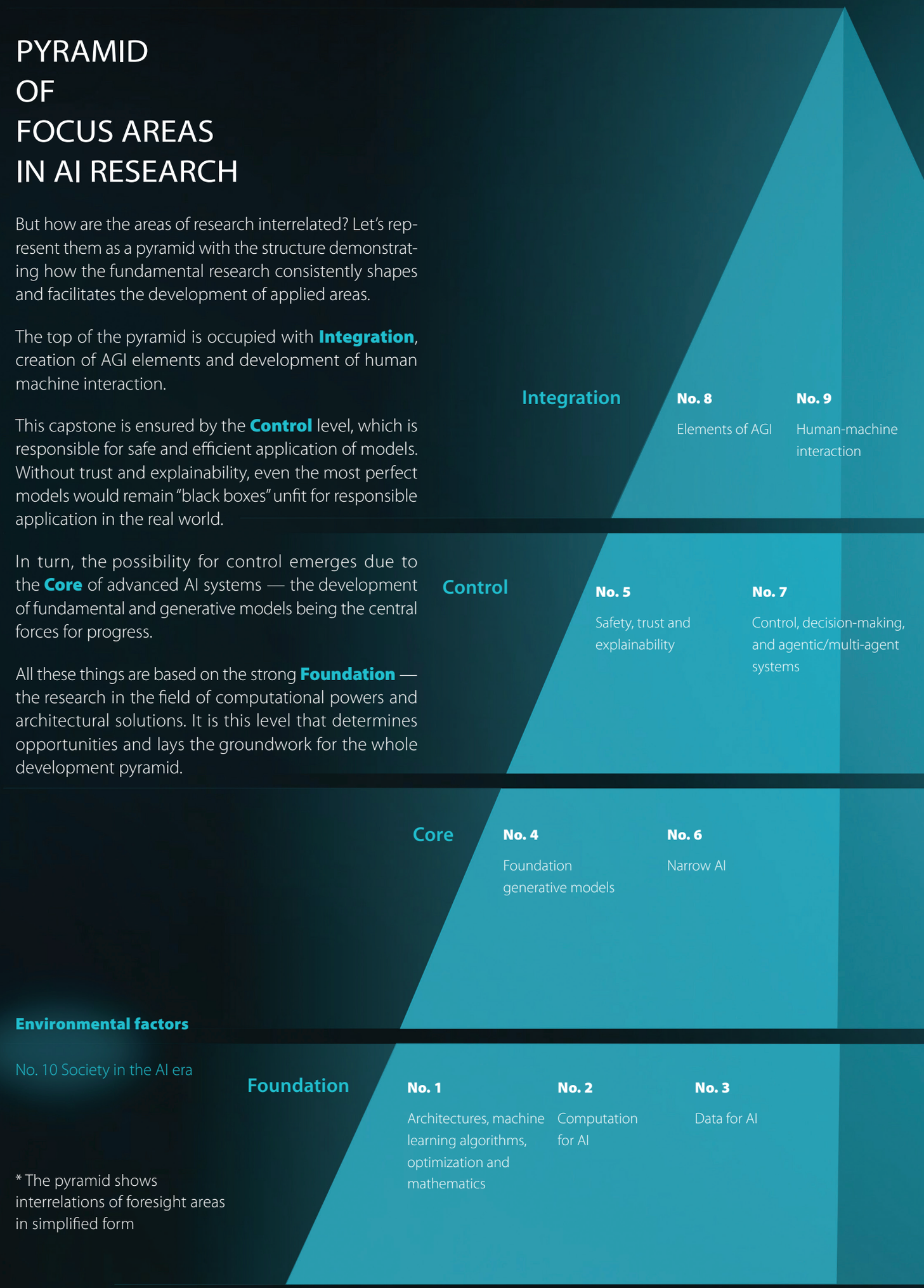
But how are the areas of research interrelated? Let's represent them as a pyramid with the structure demonstrating how the fundamental research consistently shapes and facilitates the development of applied areas.

The top of the pyramid is occupied with **Integration**, creation of AGI elements and development of human machine interaction.

This capstone is ensured by the **Control** level, which is responsible for safe and efficient application of models. Without trust and explainability, even the most perfect models would remain "black boxes" unfit for responsible application in the real world.

In turn, the possibility for control emerges due to the **Core** of advanced AI systems — the development of fundamental and generative models being the central forces for progress.

All these things are based on the strong **Foundation** — the research in the field of computational powers and architectural solutions. It is this level that determines opportunities and lays the groundwork for the whole development pyramid.




* The pyramid shows interrelations of foresight areas in simplified form

# NON-TECHNOLOGY FACTORS HAVING AN IMPACT ON THE DEVELOPMENT OF AI RESEARCH

The development of AI technologies and shaping of AI research depend not only on technological developments, but also on many non-technology factors having an impact on the pace, direction and stability of the progress. These factors form an ecosystem, where technologies are born, tested, implemented and regulated by the public opinion and government agencies. In many cases, these are these external circumstances that determine, which tasks will be treated as top priority, which data may be used, which methods turn out to be acceptable with regard to ethics, trust and legal validity, and how fast new solutions will be created and implemented.

For each focus area, the researchers identified 3–5 key non-technology factors having an impact on the area development.

- **Regulation of the technology development** frames the AI development and implementation by forcing the developers and researchers to make allowance for safety, transparency and responsibility issues.
- **The societal demand for ethics and trust** forces companies to work transparently and explain their decisions, which finally forms more reliable and safer AI systems.
- **Availability of skilled personnel** sets the pace of innovations and quality of solutions being developed and implemented.
- **The demand for cost effectiveness and replication** urges on creation of scalable and efficient solutions, which may be implemented in various industries without considerable re-engineering.
- **Availability and quality of data** have a direct impact on accuracy and generalizability of models, as well as on the opportunity for transparent risk assessment and bias prevention.
- **The growth in demand for autonomous processes** stimulates the development of systems, which are able to make decisions without constant human intervention and to retain supervision and responsibility at the same time.
- **The global demand for energy efficiency**, including the environmental agenda, promotes the creation of less resource-intensive models and more efficient data processing infrastructures.
- **The need for technological sovereignty** urges countries on independent development of critical technologies and strategic investments in their own AI ecosystems.

The expert assessments served as the basis for a consolidated visualization reflecting the landscape of key factors (Figure 1, Page 12).

The horizontal axis in the figure gives an indication of a year corresponding to the peak of trend implementation, the size of circles correlates with the weight of factors, while the colors show which focus areas are influenced by a particular factor. It should be repeated that it is the assessment of only key factors for each focus area and the absence of a certain focus area for a specific factor does not mean that this factor does not have an impact on the development of this focus area, but just implies that the impact of other factors on this focus area is assessed by the experts as more considerable.

The most important factors having the most considerable impact on the AI development are regulatory aspects, the societal demand for ethics and trust, the availability of skilled personnel, and the demand for cost effectiveness and replication.

All these factors are interrelated and form a single ecosystem: the regulations and ethics form requirements to quality and responsibility, the societal demand for trust sets the nature of user interaction, the talent pool may limit the rate of development, while the economic scalability serves as a force for practical implementation and long-term stability of technologies.

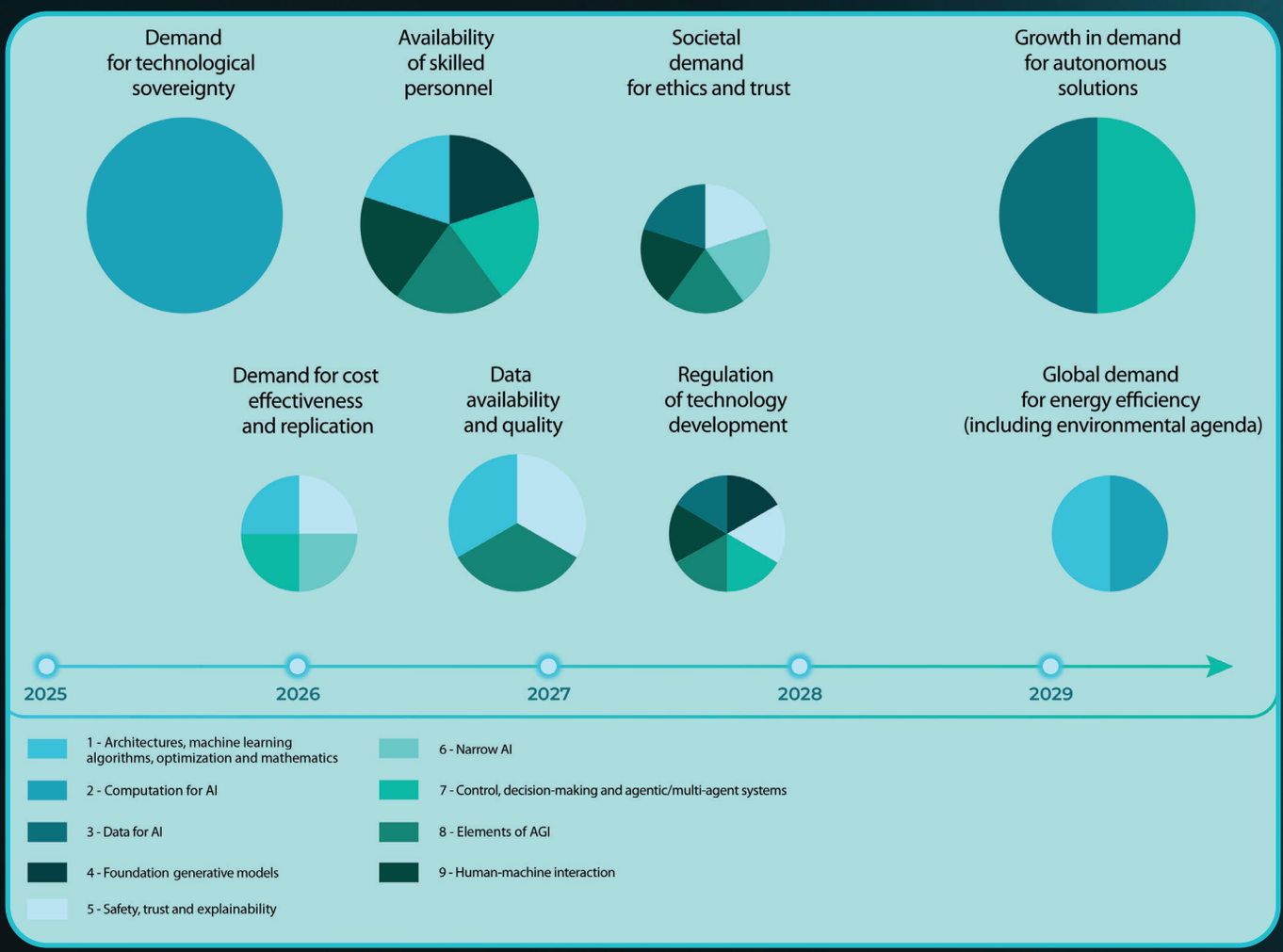

Demand for technological sovereignty
Availability of skilled personnel
Societal demand for ethics and trust
Growth in demand for autonomous solutions
Demand for cost effectiveness and replication
Data availability and quality
Regulation of technology development
Global demand for energy efficiency (including environmental agenda)
2025
2026
2027
2028
2029
1 - Architectures, machine learning algorithms, optimization and mathematics
2 - Computation for AI
3 - Data for AI
4 - Foundation generative models
5 - Safety, trust and explainability
6 - Narrow AI
7 - Control, decision-making and agentic/multi-agent systems
8 - Elements of AGI
9 - Human-machine interaction

Figure 1. Map of key non-technology factors having an impact on the development of AI research

# STRUCTURE OF FOCUS AREAS IN THE FINAL REPORT

| | |
|---|---|
| 1. Architectures, machine learning algorithms, optimization and mathematics | **1.1** Development of new machine learning algorithms<br>**1.2** AI architectures<br>**1.3** Computation speedup<br>**1.4** Distributed and federated learning<br>**1.5** Mathematical foundations of AI |
| 2. Computation for AI | **2.1** Development of specialized computing devices for AI (quantum, photonic, neuromorphic, etc.)<br>**2.2** Development of hardware and software suites for AI<br>**2.3** Machine learning and AI frameworks |
| 3. Data for AI | **3.1** Development of benchmarks for AI<br>**3.2** Data generation, conversion and maintenance<br>**3.3** Ensuring data privacy and security |
| 4. Foundation generative models | **4.1** Generative foundation models for character data<br>**4.2** Generative foundation models for non-character data<br>**4.3** Multimodal generative foundation models<br>**4.4** Knowledge transfer with adaptation of a generative foundation model<br>**4.5** Augmentation of generative foundation models |
| 5. Safety, trust and explainability | **5.1** Alignment<br>**5.2** Explainability of AI<br>**5.3** Ensuring safe development and operation of AI<br>**5.4** Ensuring protection from the results of AI use for hacking purposes |
| 6. Narrow AI | **6.1** Computer vision (CV)<br>**6.2** Natural language processing (NLP)<br>**6.3** Other narrow AI technologies (S2T, RecSys, TSA, etc.) |
| 7. Control, decision-making and agentic/ multi-agent systems | **7.1** Development of reinforcement learning algorithms<br>**7.2** Agentic systems<br>**7.3** Multi-agent systems |
| 8. Elements of AGI | **8.1** Reasoning and reflection<br>**8.2** Lifelong learning<br>**8.3** Hybrid AI<br>**8.4** Embodiment<br>**8.5** Brain and mind simulation |
| 9. Human-machine interaction | **9.1** Technical means of direct interaction with the human nervous system<br>**9.2** Technical means of traditional human-machine interaction<br>**9.3** Methods and algorithms of human interaction |
| 10. Society in the AI era | **10.1** Global AI governance mechanisms, including AI regulation<br>**10.2** AI ethics<br>**10.3** Study of AI technology impacts on society |

FOCUS AREA 1

# Architectures, machine learning algorithms, optimization and mathematics

# FOCUS AREA 1

# Architectures, machine learning algorithms, optimization and mathematics

## 1. Overview of the focus area

Almost all the state-of-the-art AI (as its forerunner, conventional Data Science) leans on three pillars:

1. Architecture/model (separating hyperplane, neural network, diffusion process, etc.). It may be considered as a function, which is especially understandable through the example of neural networks, where the choice of architecture corresponds to the choice of network structure. However, architecture does not necessary serve as the only function; it may be a set of functions, for example, for various agents. The function may be random or adaptive, changing in the course of training. Therefore, architecture may be compared to a genotype whose phenotypic manifestation depends on various factors, including external ones.

2. ML algorithms/statements: formalization of problem (learning) as an optimization or saddle problem, composing the functionality (for adversarial statements, where it is necessary to find a saddle point, the term "functionality" may be a little confusing) for residual errors (risk function), introduction of regularization, restrictions, etc. The functionality shall include the data and the model. The task is to adapt the model to the data by formalizing it as an optimization problem. Trained models should predict new data on the supposition that the law of their generation is invariable. An elementary example of that is training of Siamese neural networks. In the general case, more complex statements are possible: saddle problems (GAN), game-theoretic (RLHF) and variational inequalities, as well as multi-level optimization. The key feature of that is represented by considerable restrictions on the access to the oracle. In general, only a stochastic gradient of functionality is available (or less, as in multi-armed bandit problems), which is calculated, for example, using the backpropagation method.

3. Optimization. Solving optimization problems, saddle problems or their sets requires creation of efficient distributed algorithms. In this case it is necessary to make allowance for the convergence rate of stochastic optimization and for optimization of computation process at two levels: algorithmic, which includes the development of specialized methods for particular problem classes with the use of low-rank approximations and inexact computations, and hardware one, providing for the search for new computing architectures focused on specific features of standard problems.

This area is very dynamic, and the key factors shaping the development milestones are the following:

- Use of large language models (LLM) for the automated search for architectures, which may result in a breakthrough in their creation.
- Development of self-reflecting AI systems being able to improve themselves without constant human control, which accelerates the progress.
- Growth of interest in multi-agent systems, where AI acts collectively and autonomously in complex environments.
- Breakthrough of diffusion models in generative modeling, which changed the focus of research.
- Emergence of new optimizers (Shampoo, SOAP, Muon) being more efficient than the conventional Adam and increasing the learning performance.

The evolutionary development of this focus area allowed us to make the following conclusions for the last 5 years:

- Laws of LLM scaling show that at a certain level they outperform humans in many cognitive tasks, including creative work, which requires generation of high-quality data.

Special thanks for work on this focus area to:
A. Gasnikov, N. Bacanin, H. Houssein

- In 2024–2025, AI4Science achieved the systemic level of generation of new scientific knowledge, which may result in a "chain reaction" of acceleration in science and technology.

- It is acknowledged that efficient computing architectures should include diverse agents with various roles, which resembles the social evolution of human brain and implies the development of symbiotic neural architectures.

## 2. Overview of current developments

One of the most considerable changes over the last 1-2 years was the use of LLM for automatic proposal and assessment of new architectures for neural networks, which goes beyond the conventional numerical optimization and includes semantic descriptions of architectures. It indicates the self-referential development of AI, where the AI models are used to improve their own foundations. Efforts aimed at the computation speedup are focused on the improvement of distillation, pruning and quantization methods in order to make models smaller and adapt them to various devices. There is also the active development of some advanced methods such as speculative decoding and mixture of experts.

Key factors being the drivers for this focus area include:

### Growing power consumption by AI models and environmental/economic restrictions

The main problem of AI today is power consumption. The exponential growth of power consumption by large AI models has become a significant challenge requiring the search for new principles and optimization. The development of computationally efficient methods accelerates the improvement of "distillation, pruning and quantization" methods, as well as the development of new "efficient algorithms of "numerical linear algebra" and "optimizers".

### Depletion of data needed for training

The quality of models directly depends on the amount of data available for training. For a number of areas (for example, LLM) all the available data have been already used to a great extent. Now there is a problem where to find new ones. World Models stimulate the development of models with antecedently hardwired knowledge, which allows the considerable reduction of requirements to training sample size. Shared and private data ecosystems stimulate the development of federated learning, data exchange consortia and privacy technologies allowing several parties to make contributions and benefit from diverse datasets without disclosing private data and, therefore, mitigating the problem of data insufficiency in critical areas.

## 3. Research challenges shaping and limiting the focus area development

### CHALLENGE 1.1

#### Limited generalizability and adaptivity of AI models

Limited ability of state-of-the-art AI models to be generalized to new, never-seen-before domains and problems without considerable retraining. The problem of limited generalizability and adaptivity of AI models imposes fundamental restrictions on the widespread use of AI in dynamic real scenarios (for example, in robotics, personalized medicine) requiring flexibility and quick adaptation to new domains. The present day sees the active research in the field of new theoretical tools for deep learning, "mathematical foundation of AI", including mathematics of generative models (diffusion models, flow matching), optimal control and reinforcement learning, as well as new numerical methods of linear algebra. In the future, the development of methods to increase the "model elasticity" and "continual learning" will result in creation of more universal, stable and resource efficient AI models being able to learn like a human, which will considerably extend their applicability and trust in them.

### CHALLENGE 1.2

#### Interpretability and explainability

The problem of neural networks working as "a black box" prevents fundamental breakthroughs and broad understanding of deep neural networks work, which results in creation of "black boxes" and insufficient theoretical foundations for new algorithms. In the future, the development of theoretical tools and mathematical foundation for deep learning, including combination of statistics and optimization, will result in creation of more valid, robust and predictable AI models.

### CHALLENGE 1.3

#### Quantization and verbalization of uncertainty in AI solutions

Creates risks in sensitive areas (medicine, autonomous driving), since users may make decisions based on "self-reliant" erroneous AI forecasts, which decreases the general trust in AI and, in turn, imposes considerable restrictions on safe and reliable AI interaction with humans in situations requiring caution and confirmation of decisions, especially when models "are bad in expressing their uncertainty". Today's studies of this challenge are focused on the use of internal statistics of the models (sequence probability, entropy), development of auxiliary models for prediction of errors in generation, architectural solutions, as well as methods training models in clear expression of their certainty or uncertainty using the natural language. In the future, the development of reliable methods will allow AI to "interact with humans in a more responsible manner" signaling about the need for intervention and increasing the "trust in AI" making it a more reliable assistant being able to "reject answers".

## 4. Long-term research tasks

Development of AI systems towards more considerable autonomy and complexity gives rise to new fundamental tasks in the field of their architectures, algorithms and optimization. These tasks require in-depth studies going beyond existing approaches and involve both theoretical foundations of machine learning and applied aspects of AI interaction with humans and environment. Below you can find the long-term research tasks identified on the basis of expert discussions.

### ★ TASK 1.1

#### Creation of "algebra" above architectures, problem statements (conventionally, ML algorithms/ approaches) and algorithms for solution of emerging optimization problems

It is referred to AI selection, the systematic combination of successful components and structural blocks for generation of new architectures, problem statements and optimization algorithms applicable to complex problems reduced to known elements. Each researcher generates intuitive solutions for new optimization tasks, but it is important to formalize that, for example, through the algebraic approach to creation of efficient methods based on the conventional or stochastic gradient descent with the use of specific techniques. In spite of the availability of general principles of mathematical formalization of problems and understanding of foundations of existing architectures, the main task is the formalization of these rules and development of new architectures.

The multi-agent AI is considered as a promising tool for creation of new architectures, while mathematics ensures formalization and solution of other problems. The researchers have already obtained significant results for meta-optimization, where the search for optimal algorithm is presented as a semi-definite optimization problem successfully solved for important subclasses of convex and stochastic optimization.

Such formalization may result in the significant reduction of the search for suitable architectures and setting of their hyperparameters. In general, the process of problem solution by AI will be more formalized and, therefore, it will be possible to automate it to a greater extent.

### ★ TASK 1.2

#### Development of universal and adaptive AI models with increased generalizability

The main task is to overcome the limited ability of state-of-the-art AI models for generalization to new, never-seen-before domains and problems without considerable retraining. It is necessary to develop the "elasticity" mechanisms for memorizing and forgetting in a manner similar to human brain, in order to make "the models to generalize better and to improve adaptation to new task areas". It includes learning without forgetting and increase in "elasticity" of models.

Solution of this problem may include several approaches:

- Development of new loss functions making allowance for the efficiency of training in new problems and for the preservation of previous knowledge.
- Examination and designing of "adaptive architectures of neural networks" being able to dynamically change their structure or reconfigure themselves for adaptation to new data and tasks.
- Application of continual learning and domain adaptation.
- Study of "cognitive-inspired approaches" for the increase in flexibility and adaptability of models.

- Development of "new learning methodologies", which will allow models " to learn better", for example, on a small number of examples, just as a child, when sees just a few different dogs (big, small, shaggy) forms an abstract concept of a dog in his/her consciousness and after that recognizes any other dog easily, even that of an unknown breed.

In the future, the solution of this problem will result in creation of more universal, robust and resource efficient AI models being able for "continual learning and adaptation to new domains". It will considerably extend the practical applicability of AI in dynamic real scenarios such as robotics, personalized medicine or control systems and increase the trust in systems being able to learn throughout their "life" without the need for complete reset and retraining. The possibility to train "only on 10 instances" will allow using AI in fields with limited data.

### ★ TASK 1.3

#### Integration of scientific knowledge and World Models in AI systems

The task is to overcome the "black box" problem in the AI nature through efficient incorporation of tested "scientific knowledge" (from physics, chemistry, biology) and human expert experience directly in AI architectures and algorithms. It includes the development of World Models being able to study the basic laws of physics "by themselves" without explicit programming. It is even more relevant in fields with "scarce data" such as geology or biology.

The methods used to solve this problem include:

- Development of "hybrid AI models", including physics informing or chemistry informing of neural networks for incorporation of scientific knowledge and "not only data".
- Use of "theoretically motivated design of models", "a priori knowledge", "limitations and optimization".
- Integration of theoretical models "in loss functions".
- Creation of World Models being able to "study the basic laws of physics" by themselves.
- Development of methods for automatic "segmentation and explicit automated modeling", for example, as symbolic regression.

In the future, solving of this problem will result in the emergence of "foundation models" incorporating human knowledge in physics and other sciences, which will allow "combining theory and machine learning" in order to solve problems in fields with "very scarce data". It will result in models with the increased "robustness, extrapolation, explainability and transparency", reduction of dependence from huge amounts of data and "decrease of data and resource costs". Acceleration of scientific discoveries in chemistry, biology and physics, as well as creation of AI being able to "understand" the world at a deeper level.

## 5. Important takeaways // Expert opinion

The nature of development in this focus area is today defined by the fundamental shift of paradigms caused by the exponential growth of AI complexity. It is going hand in hand with the active search for new theoretical tools and mathematical foundations for deep learning, as well as for phenomena like "benign retraining". The breakthrough in diffusion models showed how the adaptation of quite well-known techniques to a new field may result in impressive results, which stimulates the search for new physical (or rather scientific) insights.

In parallel, there is a shift in the paradigm of optimizers — from the previously predominant Adam to more efficient matrix-oriented approaches such as Shampoo, SOAP and Muon. The development of AI is becoming more and more remarkable, when large language models are used for generation of new architectures based on basic blocks, promising a breakthrough in creation of architectures. Moreover, there is the active growth of interest in multi-agent systems that are coming to the fore rapidly as a central area in the AI research.

**55%** of research tasks, according to the estimated forecast, will not be completed by 2030

FOCUS AREA 2

# Computation for AI

# FOCUS AREA 2

# Computation for AI

## 1. Overview of the focus area

As of today, creation of specialized computing technologies and infrastructures paves the way for successful evolution of state-of-the-art AI based on big models and data. As opposed to conventional high-performance computing mainly focused on the solution of computational modeling problems, the specific nature of calculations in state-of-the-art AI is related to two aspects:

1. Building upon distinctive matrix and tensor operations effectively implemented on SIMD architectures resulting in the explosive growth in the need for GPU- and TPU-based systems, as opposed to more universal MIMD architectures on common CPU.

2. Use of big data arrays as the basis of computational process, requiring not only planning and balancing of computational load, but also control over shared distribution of data and computing ensuring the best performance.

At present, the field of computing for AI covers three key subareas:

1. Development of scalable and energy-efficient conventional computing architectures, including:

- Increase in energy efficiency of parallel computing architectures for implementation of matrix operations (GPU, TPU, etc.) while retaining linear scalability by the number of cores. The goal is to ensure the growth in the size of AI models without significant increase in costs of their training and use. Moreover, new computing architectures will also be able to stimulate the development of new AI methods and models more suitable for the use of unique capabilities of these architectures, forming the virtuous cycle of mutual promotion.

- Creation of distributed computing architectures dynamically scalable for computational tasks of AI, first of all — federated learning and LLM-based multi-agent systems. This area is critical for the development of multi-agent systems and LLM as a whole.

- Improved capabilities of specialized architectures for effective work with data and communication of agents in AI tasks, including fog computing, edge AI, etc. As opposed to multi-agent systems and large language models, this task is aimed at implementation of the problem of embodied AI (creation of autonomous agent robots, etc.).

2. Creation of computing architectures for AI based on new principles, including:

- Creation of neuromorphic optoelectronic and photonic architectures (NPU) ensuring the balance of performance and energy efficiency greatly outperforming GPU and TPU.

- Creation of quantum computing architectures adapted to AI tasks, including co-design of architectures and quantum-like computational machine learning algorithms. In spite of the fact that at present there is no impressive progress in this field, the research of quantum-like algorithms promotes outlining of future computing systems.

- Development of specialized accelerators for applied ad hoc AI tasks, including hybrid computing and FPGA-based configurable systems, as well as tools that may be used to work with them.

3. Creation of mathematical foundations, development of system- and middleware for efficient implementation of AI tasks on conventional and advanced computing architectures. At the same time, this subarea also includes the following aspects:

Special thanks for work on this focus area to:
A. Boukhanovsky, A. Phan, F. Qin

- Algorithmic mechanisms for mapping and optimization of AI algorithms with allowance for specific features of particular computing architectures.
- Automated elaboration of new AI algorithms for non-standard computing architectures, as well as co-design of new computing architectures and algorithms.
- Effective control over computational processes (planning and balancing of load, data distribution in the computer memory) for characteristic machine learning algorithms and AI tasks.
- Creation of specialized tools for implementation of AI tasks, for example, compliers and low-level frameworks for TPU.
- Development of instrumental frameworks for rapid development of AI systems for high-performance and distributed computing architectures, including creation of hybrid systems (neuro-symbolic AI, multi-agent systems and large language models, etc.).
- Development of software platforms for multi-agent systems ensuring the effective cooperation and communication between several agents. Such platforms should support the dynamic task distribution, resource control and resolution of conflicts in distributed environments.
- Development of specialized hardware platforms for multi-agent systems optimized for parallel data processing and low-latency data transfer.

## 2. Overview of current developments

At present, the research landscape in the field of computing for AI is shaped by two competing factors: development of aggregated high-performance computing technologies for large AI models, on the one hand, and distributed computing technologies for implementation of multi-agent systems and specialized ad hoc AI systems, on the other. Therefore, it is related to the active examination of the following issues:

Improvement of computational efficiency while working with large models on existing architectures, which is ensured in two stages: First, due to high-performance learning and reference methods (teacher–student distillation for model compaction, pruning, quantization, speculative decoding, as well as the use of Mixture of Experts (MoE) for computation speedup). Second, through effective mapping of resulting algorithms on the GPU/TPU architecture (which is easy to see in DeepSeek models).

Diversification of computing architectures for AI tasks, including those going beyond conventional solutions. At present, much attention is paid to neurographic spiking neural networks, optical neural networks, as well as neuromorphic computers and accelerators implementing them, which may be based on standard silicon technologies or completely alternative.

Efficient distributed computing for LLM

Training of state-of-the-art LLM requires enormous computational resources that may be available only to a distributed heterogeneous environment, on clusters with various GPU types. Control and effective use of this "GPU orchestra" remains a formidable and unsolved problem. The problem of orchestration becomes even more prominent in the light of transition to multi-agent systems of large language models.

Use of large language models in the search for architectures and adaptation of algorithms

Large language models are already traditionally used as a powerful tool for designing and optimization of neural networks as part of the Neural Architecture Search (NAS) approach. However, today's large language models can also solve the task of algorithm mapping, i.e. design neural networks with optimal implementation on specific computing architectures and choose architecture parameters for specific machine learning algorithms.

## 3. Research challenges shaping and limiting the focus area development

In spite of considerable progress, there is a number of fundamental and applied problems preventing further development and large-scale implementation of computing technologies in AI.

### CHALLENGE 2.1

Absence of drastic solution of power consumption and stability problems due to transition to new types of computing architectures

Rapidly growing power consumption of large AI models is becoming a critical restriction for their scaling. The radical reduction of power consumption requires research in the field of brand new energy-efficient computing paradigms, including quantum neural networks, holographic data representation and photonic technologies.

### CHALLENGE 2.2

**Absence of unified frameworks and problem statements for new types of computing architectures**

In spite of encouraging developments in creation of hardware basis for computing for AI, their large-scale use bumps into the absence of mature instrumental frameworks, as well as problem statements demonstrating their applicability. This gap constrains their practical application and large-scale integration. At the same time, hardware developers are typically not ready to state applied problems by themselves and often use benchmarks that are regarded as obsolete by the AI community.

### CHALLENGE 2.3

**Weak integration and joint optimization of hardware and software**

The optimal performance of AI systems requires deep optimization of combination of hardware and software components. First of all, it requires the development of specialized libraries, compilers and tools for effective connection between new hardware solutions and user applications. However, the problem remains challenging due to the entwinement of these levels.

### CHALLENGE 2.4

**Absence of understanding with regard to features of complex multi-agent AI systems in the light of use of distributed computing architectures**

Advanced MAC LLM may include hundreds and thousands of various agents, but real problems are simultaneously being solved by a much smaller number of them. Therefore, distributed computing architecture for such systems should be dynamic, with a possibility to allocate required resources for particular tasks and bind them with the optimal topology of communications. However, it requires the ability to predict the behavior of MAC LLM itself, which is still not supported by methodological framework nor even by understanding of current processes, especially in the light of collective behavior of agents resulting in emergent effects.

### CHALLENGE 2.5

**Absence of effective integration between conventional artificial neural networks and new models based on LLM**

Existing methods often use conventional neural networks as tools invoked by agents in LLM-based systems, but cannot fully integrate them and use the strengths of both. This situation results in limited capabilities in solving complex AI problems that would benefit from combined capabilities of conventional neural networks and large language models.

## 4. Long-term research tasks

In spite of general significance of long-term tasks related to creation and hardware embodiment of new computing architectures (such as quantum computing devices or neuromorphic processors), the center of gravity in their solution lies outside the existing community of AI professionals.

However, the problems related to creation of new models and algorithms specifying requirements to advanced hardware, development of tools for the mainstream use of existing solutions, as well as application for solution of applied problems may be even more significant. Therefore, long-term tasks include the following:

### ★ TASK 2.1

**Creation of processors on neuromorphic principles and their algorithms for AI purposes; sensors, environment and agents for neuromorphic processes:** with regard to creation of methodological and algorithmic AI framework for computing systems based on new principles (first of all, photonic and optoelectronic neuromorphic systems). For quantum computing devices, this task is still not very critical, due to the fact that the adequate hardware may emerge only in the remote future. This may include the examination of various issues related to adaptation of existing ML methods and AI models (for example, with regard to time series forecasting or computer vision) and to statement of brand new problems.

### ★ TASK 2.2

**Machine learning and AI frameworks**

For existing computing architectures. In this context, the most sought-after components will include frameworks for resource-intensive fields related to other foresight areas, including LLM, MAC and AGI elements, federated learning, etc. Including:

- Frameworks for symbolic and hybrid AI for creation of hybrid AI systems efficiently combining LLM and standard technologies of work with knowledge.
- Frameworks for agent-based schemes and applications (including Embodied Agents) for development of distributed systems based on heterogeneous intelligent agents.
- Frameworks for prompt engineering for fine-tuning, adaptation and customization of LLM.

### ★ TASK 2.3

**Creation of frameworks for learning and inference**

In spite of a great number of general methods for acceleration of learning and inference, their choice is completely determined by specific features of computing architecture. That's why it is essential to create instrumental frameworks being able to ensure the adaptation of set model classes to specific computing devices, based on energy efficiency and performance criteria. At the same time, such frameworks are the most sought-after in fields related to the use of special computing devices with limited characteristics (for example, on board of an autonomous robot).

### ★ TASK 2.4

**Creation of systemware improving the efficiency of work with equipment**

Intensive development of TPU results in the need for creation of the respective ecosystem of instrumental software similar to GPU, as well as tools for porting from GPU to TPU. This includes proper algorithms of tensor compilers, intermediate languages, tools for work with memory, as well as ensuring the cross-platform approach.

### ★ TASK 2.5

Distributed computing and LLM with regard to creation of methods, algorithms and software tools for control of dynamic distributed architectures for federated learning, LLM and MAC LLM. It includes the solution of problems related to ensuring multi-level parallelism in models and data for use on distributed architectures, related optimization of computing and memory usage, effective control of communications (including designing of system topology for specific tasks), as well as ensuring stability and reliability of such systems as a whole.

### ★ TASK 2.6

**Development of miniature devices being able to support large models and agents for intelligent decision-making in complex environments**

Existing large models usually require bulky devices for deployment, which imposes restrictions on their applicability in scenarios, where compactness and mobility are important — for example, in robotics. There is a need for development of miniature, but powerful devices being able to satisfy the computing demands of large models and agents, allowing robots to make reasonable decisions in complex and dynamic environments.

## 5. Important takeaways // Expert opinion

The important takeaways in this focus area include the following:

- Computing architectures for AI do not represent an independent area, since their development is inspired by creation of new methods and AI models, and by the statement of applied problems. Mathematical foundations of AI, as well as invention of new algorithms and architectures still remain very important and will move the progress forward.
- The future of AI is connected with hybrid models combining strengths of machine learning with symbolic reasoning and the use of a priori scientific knowledge. As a consequence, it requires the creation of hybrid computing systems for AI, combining specific features of both standard and new computing architectures going beyond universal computing and optimized for specific AI tasks.
- The current line of LLM development makes the growth of interest in distributed and federated

learning inevitable: in the technology itself, for adaptation and fine-tuning of LLM, and in a lighter alternative. It is indicative of the long-term relevance of such technologies, especially for AI applications on personal devices and in sensitive sectors.

- As opposed to standard high-performance computing algorithms not having their own mechanisms for mapping, planning and balancing of computational load a priori, AI provides an opportunity to build not only self-learning, but also self-adapting systems being able to reorganize their work based on requirements to performance and energy efficiency on specific architectures. This opens great opportunities for creation of embodied AI technologies implemented through co-design of proper algorithms, computing infrastructures for their execution, as well as other embodiment tools (sensors, actuators, etc.).
- Development of the ecosystem of computing for AI is sensitive not only to the access to hardware platforms, but also to the instrumental software for their use. In this context, there is a tension between advantages of open-source solutions and the strategic need for national proprietary technologies (hardware and software ones) ensuring technological self-sufficiency.

Cross-cutting, but not independent focus area significantly shaped by the development of other focus areas

**64%** of the tasks do not have a clearly defined end point and will remain relevant for many years to come

**79%** of tasks imply scientific achievements occurring in the near future (1-2 years). Scientific achievements are usually of periodic and repeated nature

The focus area development is an important and cross-cutting priority, groundwork in the field of computing for AI lays out the strong foundation for development of AI as a whole

Some computing architectures are being developed on the basis of brand new principles. The future of AI is connected with creation of hybrid computing systems combining the specific features of both standard and new computing architectures for AI

Development of new architectures is constrained by the development of systemware, which requires the simultaneous support for both areas

FOCUS AREA 3

# Data for AI

# FOCUS AREA 3

# Data for AI

## 1. Overview of the focus area

Data is the "digital blood" of state-of-the-art AI. It is needed for both training and quality assessment of AI models. According to the foresight participants, it is data, not models that are becoming the main factor of success in creation of successful AI applications. At the same time, there is a permanent shortage of high-quality data for training. It is not a coincidence that the NeurIPS conference (A* in AI) has a special track for new datasets and benchmarks every year.

At present, there are three research areas in this field.

1. Development of benchmarks for AI

   It implies creation of standardized frameworks, datasets and metrics used for measuring and comparison of AI model performance in various tasks. Each AI research task corresponds to a particular benchmark. Moreover, new benchmarks in state-of-the-art AI are the main way to set new research problems. In addition, they provide an opportunity for unbiased assessment of progress in AI solutions. As the foresight participants noted: "First of all, it is necessary to develop standards creating uniform approaches to testing of AI technologies". Research in the field of benchmarks include formation of datasets, development of methods and metrics for comparison of models, as well as creation of methods for testing of models on data emerging in real time.

2. Data generation, conversion and maintenance

   Includes a wide range of data management tasks. Data generation and augmentation are needed in such cases when the collection of required amounts of real data is very expensive, or the acquisition of real data is difficult inherently. In this context, the focus should be on approaches that ensure the integrity and reliability of synthetically generated data. In addition to generation and augmentation, the relevant tasks for this focus area are: active learning with the use of crowdsourcing, creation of simulation environments for real-time learning and testing, as well as quality assessment, filtration and data handling throughout the lifecycle of AI systems, including detection and correction of offsets and biases in data.

3. Ensuring data privacy and security

   It is a critical and always relevant research area, which allows developing AI in accordance with legal requirements and ethical standards. The research in the use of technologies concealing or falsifying sensitive information are very relevant for this area. Moreover, it is essential to use the capabilities of synthetic data in order to reduce our dependence on real data that may and must remain confidential. At the same time, detection of synthetic data and analysis of their impact on training of models represents one more top-priority task. In general, it is necessary to detect and label various types of artificial or illegally used content.

The history of AI development shows that a driver for emergence of new research areas, new tasks, an indicator of new challenges in machine learning has always been represented by creation of new iconic benchmarks such as MNIST, ImageNet, GLUE… However, as time goes by, benchmarks lose their relevance. As it was noted during one of discussions, *"we still use obsolete tests to assess new-generation models".* Evolution of AI constantly results in the revision of approaches to data handling, and over recent years, the focus area was mainly affected by:

- the development of deepfake technology and the public reaction to that;
- legal and ethical restrictions, including licenses, privacy and users' consents to personal data processing, which create barriers for collection of big data arrays;
- detection of AI models tendency towards hallucinations and offsets, which resulted in the need for reinterpretation of the data reliability concept;

Special thanks for work on this focus area to:
Y. Vizilter, X. Lin, Ch. Chen

- the development of robotics and unmanned transport, which set the problems of training in the open world and training on data generated in real time.

## 2. Overview of current developments

Today, data is regarded not only as a technical resource, but also as a political, economic and even ethical value. Major companies own unique datasets, which creates information asymmetry, on the one hand, between market entities and academia, and, on the other hand, between countries leading the field of AI development and other countries that do not have such extensive resources. This resulted in a significant trend towards creation of generally accessible datasets under public or international control.

Shortage of high-quality examples of natural text and other types of real data has now become one of the main barriers on the way to the further improvement of AI models' functionality: *"Everyone knows that we don't have enough data. Especially texts: we often just don't have natural texts. That's why many state-of-the-art models use synthetic data".* The research for creation of technologies for automatic annotation of large amounts of real data turned out to be sought-after in many fields. At the same time, creation of completely synthetic datasets turned into a key research area, especially in specialized domains, such as medicine or science: Participants of the international foresight have also mentioned the political and cultural significance of this area: *"Synthetic data are absolutely necessary for most of our local languages, since we don't have enough texts, especially in science and other fields".* At the same time, the special attention is paid to validation of synthetic data. In the course of discussions, it was noted that "if we don't control generative models, they can produce artifacts that may be easily accepted as a fact". The active research is conducted in the field of ensuring privacy and data protection with the use of such methods as differential privacy, distributed and federated learning. These methods allow working with noisy data, though models sometimes deteriorate, which remains an open problem.

## 3. Research challenges shaping and limiting the focus area development

The rapid development of AI reveals bot fundamental and applied challenges in data handling.

### CHALLENGE 3.1

#### Data reliability and representativeness

As it was mentioned by the foresight participants, *"the reliability of data is the basis for the reliability of models".* The community is more and more concerned with the data reliability. The special concern is caused by AI hallucinations — generation of non-actual or false data. There are some problems with representativeness: most datasets are focused on the English language and western culture, which causes data offsets and makes models less efficient in other regions. In this context, as one of experts emphasized, *"data is a form of political power".*

The impact of this challenge is related to the fact that reliable data is the basis for trust to AI conclusions and decisions. Unreliable data may result in inaccurate forecasts, biased models and expensive failures in AI usage. Perhaps, the most important aspect here is ethical AI: elimination of data bias guarantees validity and impartiality of AI arguments.

The data reliability is ensured through the use, research and development of such tools as data management methods, data quality assurance methods (strict data collection protocols, continuous data check and cleaning), methods for detection and elimination of hallucinations, implicit and explicit bias in datasets. Creating and using one's own reference datasets represents an important tool for combating bias.

### CHALLENGE 3.2

#### The need for high-quality synthetic data

One of the central challenges of modern AI is the so-called "data wall" problem. Laws of model scaling showed that the improvement of their work requires the increase in the amount of data used in training. However, many fields already train models using all the available data from the "digital footprint" of mankind.

The scenario of further development of AI is directly related to the solution of this problem: whether it is possible to improve the quality of models' work further, or the maximum level of AI development is already reached, since it is limited to the available amount of data produced by the mankind. From the practical point of view, the question is how we can go beyond the initial training sample. From the fundamental point of view, the challenge is almost philosophical: can a teacher

teach a pupil who will become superior to himself? In other words, are we going to have the superhuman AI in the foreseeable future, or is there an essential prohibition on its emergence?

For the last year and a half, the community efforts were focused on solution of the "data wall" problem through generation of synthetic data, though the proper synthesis of new data being in line with distribution of real data still represents a complex and relevant research task. The considerable progress is achieved in such fields as reasoning, programming and mathematics, where it is possible to check the quality of generations in unbiased manner. As it was mentioned by the foresight participants, assessment of actual accuracy of AI generated content in other fields is very labor-intensive, while the absence of reliable metrics makes the unbiased assessment of reliability and quality of output data a very difficult task. Evaluation beyond a person's subjective opinion remains an unsolved problem: *"Most often, we don't have an objective mechanism. We can rely on people's opinions, but then we are limited by human faculties".*

In the field of robotics, Embodied AI, engineering and business applications, simulation environments are playing an increasingly important role as a dedicated data source for modeling dynamic processes and the operation of LLM/VLM/VLA models in real time. Both physical simulators and business process models, calculation and engineering software are used. The main challenge, however, is to ensure that such digital twins are both realistic and fast. Speed is required to perform multiple cycles of reinforcement learning. Realism is important for transferring learning from virtual to real-world environments. As the foresight participants noted, "transferring models from simulators to real-world conditions remain a serious bottleneck. Bridging the sim2real gap is key to the adoption of AI models".

### ⚡ CHALLENGE 3.3

#### Ensuring data privacy and security

As the foresight participants noted: *"Data privacy is definitely the most important area right now. More and more data are being created, more and more privacy violations are occurring. This work to strengthen data protection measures must be ongoing, but breakthroughs are possible within a few years".*

Data privacy and security in AI is a prime example of an area where AI technologies simultaneously create threats and provide means of protection against them. If AI and machine learning tools can provide data privacy guarantees, this will enable regulatory compliance and build user trust, particularly in areas such as healthcare, finance, and e-commerce, where data privacy is critical.

The foresight participants note that the integration of differential privacy (DP) and leakage testing methods should become standard in the development of training pipelines. Technologies like these help comply with privacy laws, such as GDPR and standards like ISO/IEC 42001, by providing quantitative privacy guarantees. At the same time, it is necessary to take into account and investigate various schemes of attacks on data privacy in machine learning models. For example, when using so-called Membership Inference Attacks (MIA), an attacker observes the output of a trained model and tries to determine whether a certain data record was part of the training set. In this context, a pressing research challenge is not only to combat such attacks based on DP and other methods, but also to competitively generate new attacks, which will facilitate the further development of more advanced protection systems.

Generating synthetic data based on private real data can also be a way to protect their privacy, albeit at the cost of some loss of data uniqueness. This is especially important in sensitive areas such as medicine.

## 4. Long-term research tasks

### ★ TASK 3.1

#### Data management, data quality assessment, and filtering, curation, and sorting methods

Data management serves as a critical foundation for ensuring the quality, security, and usability of data in the development of modern artificial intelligence. As data scales rapidly expand and AI application scenarios become increasingly complex, AI projects that lack systematic data management often face numerous challenges, such as data inconsistency, labeling errors, sensitive information leaks, and compliance risks. Effective data management not only ensures high-quality, standardized, and traceable data sources for training models, but also establishes clear accountability mechanisms and quality control standards throughout the lifecycle of data creation, labeling, integration, and use. Data management, especially in multi-source, heterogeneous, and large-scale data environments, significantly improves data reliability, reusability,

and ethical compliance through techniques such as metadata management, provenance tracking, access control, and quality monitoring.

Data quality assessment (DQA) is performed according to the criteria of accuracy, completeness, consistency, reliability and uniqueness.

Data filtering involves removing irrelevant, erroneous, or noisy information. Furthermore, data filtering has been a powerful tool since the days of the paper titled "Textbooks Are All You Need" to overcome scaling laws, allowing for significant reductions in data volumes, training time, and computational budget with the same achievable level of results practically reduced to a minimum.

### ★ TASK 3.2

#### Data privacy in federated learning technologies

Federated learning (FL) is a specialized subset of the broader distributed learning paradigm that involves training a model simultaneously on multiple servers without the need for data centralization or exchange of source data. Local servers train local copies of the model and only transmit information needed to update the model (such as gradients or weights) to the central server. The general model is updated, after which copies of it are sent back to the local servers, and the training cycle is repeated.

Storing data on local devices ensures data privacy and compliance with data protection regulations. This also reduces the risk of data breaches, as the creation of large centralized datasets always comes with the potential for greater vulnerability.

Federated learning enables multiple organizations with their own private data banks to participate in the creation of innovative AI solutions. At the same time, the risks of disclosure of confidential information for them are significantly reduced, and with the simultaneous use of additional protection mechanisms such as DP, they can be practically reduced to a minimum. This is critical in industries such as telecommunications, healthcare, and manufacturing, where data privacy and regulatory compliance are critical.

### ★ TASK 3.3

#### Data privacy in federated learning technologies

Detection and separation of artificially generated content has historically had several aspects: detection and labeling of falsified or artificially created media files, analysis of text information to identify its artificial origin, and determination of whether there is synthetic information in training samples.

Deepfake detection has traditionally relied on searching for evidence of data manipulation, such as visual artifacts, timing irregularities, and unique camera or microphone characteristics. However, there is an increasing need to identify the "authorial style" of generative models that do not leave obvious inconsistencies in the generated data.

Detection of generated text using NLP methods can rely on statistical analysis, linguistic features, and informational characteristics of the text, such as complexity and perplexity. Recently, the characteristics of texts created by LLMs and humans have become increasingly difficult to distinguish, so it is necessary to find new ways to identify synthetic texts based on machine learning methods.

Watermarking of synthetic content involves the intentional embedding certain invisible artificial patterns in the output of generative AI models, allowing them to be subsequently detected with high accuracy.

In this case, it is possible to either create data and watermarks together, or to create them sequentially. The first option is more promising, since the generated data can immediately contain watermark information, but the second approach is still used more often.

Addressing this range of issues is crucial to creating effective means of combating disinformation, which can have serious social consequences. Deepfake technologies are also associated with economic fraud (phishing, fake calls and video messages), and detection technologies are designed to help protect companies and individuals from financial losses, data breaches and reputational damage. There are also a number of areas (such as education) where it is important to confirm that the text was written by a human (for instance, in student essays).

The problem of training data contamination stems from the fact that training data are often extracted from the Internet. The foresight participants have repeatedly noted: *"It is crucial to be able to identify synthetic data and evaluate its impact on the learning process".*

## 5. Important takeaways // Expert opinion

The AI research community emphasizes the need for open collaboration, a broadened research agenda

in data science, and a strong focus on ethical and regulatory considerations for ensuring the responsible development and implementation of AI technologies.

Most participants agreed that data quality is a major constraint on AI development. Model reliability is based on data reliability. Data privacy is a critical and extremely topical area of research. Creating and using one's own reference datasets is essential for conducting research and combating bias. The focus should be on approaches that ensure the integrity and reliability of synthetically generated data. It is crucial to be able to detect synthetic data not only to prevent threats like deepfakes, but also to assess their impact on the learning process.

There is consensus that cooperation and data sharing accelerate development and are particularly important for countries with limited digital resources that are seeking to develop their own AI models. However, the responsible use of AI technologies requires formalized mechanisms such as certification.

International standardization and certification of benchmarking methods (testing and evaluating AI systems) is crucial for mutual recognition of AI developments and effective communication in this field. At the same time, soft regulatory frameworks for AI technologies are preferable to strict legislation: *"We need to start by creating a regulatory framework… Perhaps not laws in the strict sense, but a framework that will allow for streamlining interactions with AI".*

Benchmarks in modern AI are the main way to set new research problems

**47%** of the tasks in the focus area are directly related to the issues of generation, transformation and maintenance of data

**30%** are related to ensuring data security and confidentiality

Data quality is a major constraint on AI development, and the reliability of data is the basis for the reliability of models

A critical challenge for modern AI is the "data wall" problem. Is further improvement possible or has the maximum level of AI development already been reached, as it is limited by the available amount of data produced by humanity?

Responsible development of the focus area requires open collaboration, an expanded research agenda in working with data, and special attention to ethical and regulatory aspects

FOCUS AREA 4

# Foundation generative models

# FOCUS AREA 4

# Foundation generative models

## 1. Overview of the focus area

The field of foundation and generative models is characterized by rapid development, high requirements to computational resources, and rapidly changing research priorities. Key trends include the development of generative model capabilities to expand their application areas, as well as the increasing role of synthetic data for training models. This shift is driven by a lack of new inputs and the desire to surpass human performance levels.

Currently, special attention is paid to the following sub-areas of foundation and generative models:

### 1. Foundation generative models for character data

Foundation and generative models for character data underlie the modern development of large language models (LLMs). The main focus area is related to the development of systems capable of meaningful generation and interpretation of complex character structures, including logical reasoning, program code and formalized knowledge. Research is underway to improve factual accuracy using mechanisms such as Retrieval Augmented Generation (RAG) and to optimize performance through accelerated inference techniques (e.g., speculative decoding).

### 2. Foundation generative models for non-character data

Diffusion models and neural field-based methods, including NeRF and Gaussian Splatting, play a key role. They enable the simulation of complex physical and visual processes, creating lifelike images, 3D scenes, and animations. At the same time, the co-optimization of algorithms and hardware — from sensors to processors — is developing.

### 3. Multimodal foundation generative models

The development of multimodal models is critical to the creation of anthropomorphic robots capable of natural interaction with people and the environment. These models enable robots to simultaneously process and interpret data from various sensors (vision, hearing, tactile sensors), which is necessary to perform complex tasks in dynamic real-world conditions. Without such technologies, it is impossible to achieve full-fledged social interaction, autonomous decision-making, and adaptive behavior in unpredictable situations. Multimodal models are thus a key element in bridging the gap between limited automated systems and the robots of the future.

### 4. Knowledge transfer and adaptation of foundation generative models

An important area of development is the adaptation and transfer of knowledge from foundation models to new domains and tasks. AI scientists are investigating methods of additional training and parametric adaptation (fine-tuning, LoRA, mixture of experts), ensuring efficient use of already trained models in data-poor areas. Approaches to the transfer of knowledge between modalities and domains are developing, including in the context of continual learning.

### 5. Augmentation of foundation generative models

Modern research is also aimed at expanding the capabilities of generative models themselves through augmentation mechanisms. This includes the connection of external memory and tools (tool use, memory augmentation), the use of external knowledge bases and simulators, as well as the creation of systems capable of self-diagnosis and self-learning. Augmentation also encompasses the use of synthetic data to expand training sets and improve the robustness of models.

The modern era of large language model development began in 2017, marked by the emergence of the transformer architecture and the attention mechanism. This revolutionary innovation served as a catalyst for rapid progress in the field of text generation and processing.

Special thanks for work on this focus area to:
A. Kuznetsov, H. Haowen, Y. Tian

- **ChatGPT Breakthrough (2022):** the release of ChatGPT marked a shift in large language models. The emergence of ChatGPT and its analogues (2022) has become a turning point: the explosive growth of interest in generative models shifted the focus from narrow NLP tasks to general-purpose conversational systems.
- **The multimodality of LLMs:** the addition of visual, audio, and video modalities — has expanded their applicability: models have learned to recognize images, analyze videos, and reason about complex questions, solving a wide range of problems.
- **Embodied AI** has enabled the application of multimodal models in robotics and autonomous systems (Vision Language Action), paving the way for more intelligent agents.
- **Diffusion models** have become the basis for image and video generation, surpassing GANs. Models like Stable Diffusion and Sora set a new standard of quality and stimulated the development of areas such as synthetic data generation and RL tuning of visual models.
- **The growth of open source initiatives** (LLaMA and others) has democratized access to advanced technologies, allowing the creation of national and domain-specific models adapted to linguistic and cultural characteristics.

## 2. Overview of current developments

The modern landscape of foundation and generative models is dynamic. It is distinguished by a desire for universal possibilities, the growth of multimodality and integration of models into a wide range of applications.

### Large language models (LLMs)

This type of AI models is increasingly being used for text processing and user interaction, including integration with hardware systems to perform real-world actions. Methods are being developed to expand their capabilities without complete retraining, in particular Retrieval Augmented Generation (RAG), which ensures the relevance and accuracy of knowledge. In parallel, research is being conducted to optimize prompt engineering and accelerate inference (speculative decoding, Mixture of Experts), which increases the efficiency and scalability of systems.

### Multimodality as a key direction

The development of multimodal models capable of processing and generating data in different modalities one of the leading trends. This includes technologies for understanding images, audio and video, processing long context, understanding spatial and spatiotemporal relationships of objects, and generating images and video, which are often integrated into a single architecture.

### Diffusion and other models for multimedia data

Multimodal generative models are rapidly advancing, demonstrating increasing ability to synthesize content based on different modalities, such as generating high-quality images and videos from text descriptions. Diffusion models have become the cornerstone of this progress, providing unprecedented quality and control over multimedia data generation. Their applications extend beyond the creative and media fields — they are used in robotics to plan actions based on perception, simulation, and environmental modeling. These advances represent a transformational leap in AI's ability to interpret and reproduce complex real-world sensory data.

### Multimodal models and anthropomorphic robots

Key challenges in the development of humanoid robots are the creation of optimal multimodal sensor systems with efficient data fusion and ensuring real-time information processing with limited computing resources. In parallel, urgent tasks remain, including overcoming the shortage of specialized data, achieving socially acceptable interactions without the uncanny valley effect, and ensuring the ability to generalize and adapt in diverse environments without large-scale retraining.

Current approaches to modeling the physics and geometry of the world face fundamental challenges, including the violation of physical laws by data-driven models, limited generalization ability for non-distributional data, and insufficient scalability for complex environments. Critical areas of development include integrating physical priors into model architectures, ensuring robustness to unknown scenarios, creating efficient computational methods for large-scale modeling, and achieving interpretability and cross-modal causal reasoning. Integration with search and hybrid approaches the use of hybrid approaches that combine generative models with search algorithms

and external tools to improve the quality of reasoning and data generation is growing. Evolutionary methods like AlphaEvolve are becoming a promising focus area. They combine language models with evolutionary algorithms to automatically generate and improve code, allowing for the creation of complex, interpretable solutions without human intervention.

## 3. Research challenges that stimulate or limit the focus area development

### CHALLENGE 4.1

#### Hallucinations of foundation and generative models

"Hallucinations" remain one of the main challenges — models generate plausible but incorrect information, which undermines trust and limits the use of AI in critical areas such as medicine and public administration. Overcoming this problem requires the development of objective methods for assessing quality and reducing the number of errors. Approaches that increase reliability are being actively researched worldwide, including Retrieval Augmented Generation (RAG), as well as hybrid and neurosymbolic methods aimed at creating transparent and trusted AI systems.

### CHALLENGE 4.2

#### Computational cost and efficiency

Training and running large models requires enormous computational resources and energy, which is becoming a major barrier to scaling and mass adoption of AI. High costs limit access to technology and slow down progress, especially in tasks that require real-time operation. In response to this challenge, energy-efficient architectures (e.g., Mixture of Experts), optimization methods (quantization, distillation), specialized chips (TPU, NPU), and distributed computing systems are being developed to reduce costs and increase the availability of AI.

### CHALLENGE 4.3

#### Generalization and transfer of knowledge

Despite success in solving individual problems, the models still have poor generalization skills and perform poorly in new conditions not encountered during training. Limited generalization hinders the creation of universal and autonomous systems that can adapt without retraining. To overcome this gap, transfer learning and domain adaptation methods, hybrid and evolutionary approaches such as AlphaEvolve, and the integration of prior knowledge, including physical laws, are being developed to improve the robustness and adaptability of models.

## 4. Long-term research tasks

### ★ TASK 4.1

#### Creating computationally efficient architectures for foundation generative models

Developing methods that allow training models of comparable complexity with significantly smaller amounts of data and computations. The main problem with modern systems is the high resource intensity of training; the challenge is to improve its efficiency without sacrificing quality. The solution involves the use of faster optimizers, knowledge distillation, synthetic data generation, Mixture-of-Experts architectures, and continuous learning. Co-design of algorithms and hardware for maximum performance remains an important area. This solution will enable training models of the same level using tens of times fewer resources, making the development of foundation models available to a wide range of research groups.

### ★ TASK 4.2

#### Research and development of methods for creating generative models (including RL in various applications)

Modern foundation models have reached the limit of the amount of available "human" data, so the key challenge is to teach them to independently generate high-quality synthetic data and learn from it using feedback from the environment. This is ensured through the use of reinforcement learning (RL) methods, where models interact with simulations or the real world and improve reward-based strategies (e.g., DeepSeek R1), and hybrid approaches with search algorithms (MCTS) and learning on unlabeled data. Solving this problem will lead to the emergence of self-improving models capable of learning without a constant influx of new data, which will ensure a qualitative increase in their abilities for reasoning, planning, and autonomous action, including application in robotics and agentic systems.

## ★ TASK 4.3

### Development of fine-tuning methods for foundation generative models (e.g., LoRA, P-tuning)

Foundation models contain extensive knowledge, but adapting them to new domains (e.g., medicine or law) requires expensive retraining and often leads to the loss of previously learned information. It is necessary to create effective methods for transferring and integrating knowledge without complete retraining. For this purpose, adapters and LoRA are used, which allow training individual modules without changing the main model. Moreover, approaches such as RAG are put in place, which use external knowledge bases. Distillation of knowledge to transfer information into compact models and "unlearning" methods to remove unwanted data are promising. The result will be the development of flexible and cost-effective adaptation technologies that enable the creation of accurate and specialized domain and national models while complying with data protection requirements.

## 5. Important takeaways // Expert opinion

This focus area can be characterized as one of the driving forces of modern AI: to achieving this are quickly being translated into all other fields and generating new approaches to solving old problems.

The expert community agrees that the future development of foundation and generative models will be determined by hybrid approaches. The greatest progress is expected to come from combining machine learning with symbolic reasoning, and from creating multi-agent systems that can interact in a complex manner with each other and with the environment.

Despite the impressive power of large foundation models, the critical need for their adaptation and retraining (localization) is recognized. This applies not only to the support of low-resource languages and the consideration of cultural contexts, but also to applications in specialized fields such as medicine or engineering, where specific knowledge is required.

Standardized and open benchmarks are considered the basis for sustainable progress. They are necessary for an objective evaluation of models, ensuring the reproducibility of studies and reliable comparison of results. Overcoming the lack of reliable metrics, especially for evaluating generative texts, remains a significant challenge. The problem of assessing creative tasks without a correct answer option should also be highlighted separately. How can we determine that one text is better than another, or one image is better than another. To solve this problem, scientists are actively developing the training of critic models (large language or multimodal models) that act as judges in evaluating the generated content.

A rapidly growing area, the potential of which for applied applications is difficult to overestimate

The most important trend in the development of the focus area is multimodality; the importance of research here will grow significantly in the coming years

**34%** of the tasks in this area are directly related to text and character data,

and another **41%** — are indirectly related to such data

The most important breakthrough will be the creation of models that understand the physics and geometry of the world; this will open up a whole range of new possibilities and applications and can affect the development of almost all other areas

One of the most serious challenges for generative models remains the problem of "hallucinations" — the generation of false or unsubstantiated information

FOCUS AREA 5

# Safety, trust and explainability

# FOCUS AREA 5

# Safety, trust and explainability

## 1. Overview of the focus area

The Safety, Trust and Explainability of AI area covers a range of tasks related to the development and operation of intelligent systems, ensuring their reliability, predictability, and social acceptability. It defines the boundaries of the applicability of AI technologies: it is the degree of trust on the part of society, the government, and business that determines the possibility of their integration into critical areas.

In contrast to the traditional understanding of security, which was limited only to the technical soundness of software, the modern approach integrates issues of the stability of models, their consistency with human values, and the ability to provide transparent, verifiable, and reproducible grounds for decisions. Research in this area is interdisciplinary, combining methods from computer science, mathematics, cybersecurity, law, ethics, and social sciences. Viewing AI development through the prism of trust makes it possible not only to develop practical solutions but also to formulate new fundamental AI problems that require the development of a consistent theory and engineering practice for its safe operation.

Within this focus area, the following subareas are distinguished:

### 1. Alignment

The goal of alignment is to prevent AI models from generating harmful results — outputs that are unreliable, contrary to societal values, violate laws, or are aimed at illegal activities. To achieve this, AI scientists are developing methods for fine-tuning models, introducing additional levels of control, testing, and benchmarks, as well as mechanisms for forming value systems in generative models for widespread use.

### 2. Explainability of AI technologies (XAI)

Explainability ensures transparency and trust when using AI in critical areas such as medicine, governance, or law. Research includes the development of methods for post factum explaining "black boxes", the creation of transparent models based on logic and knowledge bases, and the formation of requirements and protocols for testing AI systems for explainability.

### 3. Ensuring safe development and operation of AI technologies

To integrate AI into critical systems, it is necessary to develop a consistent theoretical and technological framework that ensures trust in intelligent systems.

This should include not only checking the program code and libraries used for vulnerabilities, but also identifying and eliminating defects in datasets and models specific to AI. Key areas include protecting data and models from attacks, developing secure development methods (MLSecOps), and creating specialized tools and benchmarks to assess the level of security and trust in AI systems.

### 4. Ensuring protection against the results of using AI for the purpose of hacking

AI can be used for hacking, cyberattacks, social engineering, and information forgery. Important areas of research include identifying vulnerabilities and deepfakes, developing data protection methods, including digital watermarking technologies, and creating tools to counteract the illegal use of intelligent systems.

Initially, the security of AI systems was understood solely as technical correctness and fault tolerance. The main discussions focused on defining the distinguishing features of AI within the traditional paradigm of secure software development. However, it quickly became apparent that program code analysis methods were not able to detect defects in datasets and machine learning models, which required the development of new specialized approaches and tools.

With the expansion of AI applications in socially and economically significant areas, systemic risks associated

Special thanks for work on this focus area to:
A. Avetisyan, A. Rocha, O. Rogov

with discrimination, vulnerability to attack, and a lack of predictability and transparency in decision-making have begun to emerge. This has led to a broadening of the focus of researchers and developers from checking "whether a system works technically correctly" to a broader understanding of "whether it provides security and fairness to society".

The following key milestones have formed the foundation for the development of this field:

- **The widespread adoption of generative AI** has demonstrated the technology's potential, but has also revealed problems with hallucinations, indeterminacy, and abuse (deepfakes, destructive content). The expansion of agentic capabilities (tools, memory, web access) has dramatically increased the attack surface and the frequency of model behavior policy violations.
- **Autonomous vehicle accidents** have confirmed the need to develop new methods for ensuring functional reliability in conditions of uncertainty.
- **Identified cases of discrimination** in recruitment and credit scoring systems have highlighted issues of fairness and bias in algorithms.
- **The beginning of the development of international standards and legislation (EU AI Act, IEEE initiatives)** has strengthened institutional regulation and established basic requirements for the transparency and security of systems. Evolving requirements shift the focus towards auditing, certification, and operational monitoring of AI systems.
- **Tightening data protection regulations (GDPR and similar ones in other countries)** have stimulated the development of differential privacy and federated learning methods.
- **Recent research in AI security has highlighted the key role ofadversarial robustness,** showing that protecting models from such impacts is essential for their safe deployment in critical areas.
- **The emergence of hidden distributed instructions in external data** has become one of the new central threats to autonomous agentic AI. The high effectiveness of such attacks requires fundamental and multi-layered adaptations to the policies for developing, integrating, and operating AI models.

The Safety, Trust and Explainability of AI area is critical for the sustainable implementation of AI in socially and economically significant areas. The trust of society, business, and the state in intelligent systems determines the boundaries of their application: without confidence in the reliability and fairness of AI decisions, its potential remains limited. Developing safe and explainable AI methods enables wider adoption of the technology while reducing the risks of error and abuse.

## 2. Overview of current developments

Over the past few years, the field of trusted AI has seen a shift from conceptual discussions to the implementation of practical tools. The widespread use of generative models has brought to the forefront the challenges of filtering input and output data, implementing digital watermarks for protecting against abuse, and developing methods for monitoring models in operation. In the area of explainability, architectural approaches (creation of interpretable models and methods) and methods for post-factum explanation of results to the user (visualization and analysis of decisions of already trained systems) are being developed in parallel.

Software tools for developing secure AI systems are being created and improved by analogy with tools for developing secure software, and methods for their application are being worked out. These tools include both tools for analyzing training data and tools for testing the security of trained models. Leading organizations are creating dedicated AI security teams that stress-test models by simulating malicious attacks, bypass attempts, and exploitation of hidden vulnerabilities. The goal is to identify and eliminate risks before public deployment of the model.

Tools and methods that ensure the safe operation of AI models are being developed and improved. This includes filtering the model's input and output (censoring), detecting and preventing attacks, anomalies, information leaks, and attempts to steal (unauthorized distillation) the AI model, and is complemented by incident response procedures and logging of model decisions. Both classical, non-intelligent tools and AI models specially trained to ensure the security of other AI models can be used.

Increasing attention is being paid not only to the final decision of the model, but also to the degree of

its confidence in this decision. Methods are being developed to allow the model to calibrate its predictions and signal low confidence in unfamiliar or borderline situations.

Reinforcement learning from human feedback (RLHF) has become the de facto standard for aligning large language models. Reinforcement learning from AI feedback (RLAIF) techniques, where the alignment of one AI model is implemented by another AI model, are becoming increasingly popular. Research is underway to empower AI models to evaluate the validity, safety, and ethics of their own judgments as they are generated.

In the field of explainable AI, there are two areas:

1. Interpretable models, where transparency is built into the design phase: these are models in which the decision-making process is intuitively understandable to humans. These include decision trees, linear models, and hybrid architectures with visualized internal dependencies.
2. Post factum explanation focuses on developing methods for interpreting already trained models, including complex neural networks. Approaches such as SHAP, LIME, visualization of attention mechanisms and generation of explanations for users and developers are used here.

It is important to note that full and transparent interpretability of modern models without loss of quality is practically unattainable; the goal is not complete transparency, but rather ensuring sufficient explainability and trust while maintaining high efficiency.

Both approaches complement each other: the first one ensures "out-of-the-box transparency", while the second one allows for the interpretation of complex models where built-in interpretability is limited.

Real-world testing of agentic AI reveals critical vulnerabilities: during a large-scale public red team, over 1.5 million model calls were completed, and over 60,000 successful policy violations were recorded in 44 scenarios for 22 models. Particularly effective are indirect injections through third-party downloaded data, such as web pages, documents, and emails.

## 3. Research challenges that stimulate or limit the focus area development

### CHALLENGE 5.1

#### The lack of a rigorous formal theory of machine learning

Modern AI is largely developing in the paradigm of finding new solutions through trial and error. The lack of a rigorous formal theory limits the ability to predict the boundaries of applicability of technologies, determine the conditions of their reliability and safety. Without a fundamental basis, it is impossible to answer the key question: where and under what conditions do modern methods cease to be effective and safe. In response to the challenge, research is intensifying in the fields of mathematical foundations of machine learning, the theory of generative models, and the formalization of the concepts of trust and resilience.

### CHALLENGE 5.2

#### Alignment

The main challenge is to ensure that the behavior of generative AI conforms with social values and legal norms. The threat is that models may generate unreliable, prohibited or dangerous results, or be used for illegal activities. The failure to address this issue undermines trust in AI and limits its implementation in socially significant areas. In response to this challenge, work is underway around the world to develop value alignment methods, testing systems, and benchmarks for assessing the reliability of models.

### CHALLENGE 5.3

#### Explainability and transparency of decisions

The complexity of modern neural network models hinders understanding of their internal logic. This creates risks when making important decisions in medicine, law and governance. The lack of explainability limits the practical applicability of AI and increases public skepticism. To overcome this challenge, methods for creating interpretable models and post-factum explanation methods (SHAP, LIME, etc.) are being developed, and standards and protocols for assessing the level of explainability are being formed.

### CHALLENGE 5.4

#### Secure Development and Operations (MLSecOps)

The growing use of AI increases the threats posed by the introduction of vulnerabilities into code, libraries, and datasets. Attacks on models during training and execution (dataset poisoning, adversarial attacks on

models) can lead to system failures and compromise of critical systems. In response, AI security engineering (MLSecOps) approaches are being worked out to ensure the security of the entire lifecycle of models. New methods are being developed to counter indirect injections at the tool integration stage, including call chain analysis and input data filtering. Benchmarking of resilience to agent-based attacks is actively developing using reference sets of attacks and regular reassessments of the reliability of models. At the same time, policies for secure work with untrusted sources are being developed, including the creation of isolated sandboxes, the use of lists of permitted domains, and the implementation of automated content checks before loading it into the model.

### ϟ CHALLENGE 5.5

Countering AI abuses

AI models can be used for malicious purposes, ranging from hacking and social engineering to creating deepfakes and fake information. This threatens the security of individuals, organizations and the state. The challenge requires the development of data protection methods (for example, using watermarks), counterfeit detection algorithms, and systems to counter cyberattacks.

## 4. Long-term research tasks

### ★ TASK 5.1

Developing methods to reduce the risks associated with incorrect or harmful data (Inaccurate data, restricted data, including information that is prohibited from dissemination by law, data that are not consistent with societal values)

To solve this problem, it is necessary to develop methods for forming value systems in generative models, tests and benchmarks, as well as to test the resilience of models to vulnerabilities. For agentic AI systems, alignment should extend to tool-use chains, including assessing the safety of intentions and side effects before making tool invocations. It is important to consider the impact of optimization procedures during development, as well as the stability of individual knowledge domains during such actions.

### ★ TASK 5.2

Formulating common approaches to ensure explainability and increase trust in the work of artificial intelligence — Explainable AI (XAI)

There is an emerging trend towards researching and creating transparent AI methods based on formal systems, logic, and knowledge bases. Formalization of requirements, protocols, evaluation indicators, and benchmarks for testing systems for explainability is also underway. Operational explanations for the user or auditor are also important: why the tool was called, what policy constraints were taken into account, how risks were assessed.

### ★ TASK 5.3

Creating methods and infrastructure to ensure secure development of AI systems (MLSecOps)

This task involves developing and implementing principles for building a secure development infrastructure (MLSecOps), working out methods for protecting against dataset poisoning and attacks on models, and tools for searching for vulnerabilities in third-party libraries. Benchmarks are being developed to formalize the level of trust in AI systems.

### ★ TASK 5.4

Developing methods for detecting and protecting against deepfakes, including watermarks

Developing countermeasures against deepfakes is becoming a critical task in ensuring digital security. The widespread use of generative AI has made the creation of realistic fake video and audio materials available even to untrained attackers, necessitating the urgent development of effective protection methods. Digital watermarking technologies are becoming especially relevant, allowing not only the detection of counterfeits post factum, but also the pre-marking of legitimate content, creating the basis for verifying its authenticity. These solutions are becoming very much in demand in the context of combating disinformation and protecting personal data, where the ability to quickly identify counterfeits directly impacts the protection of user rights and freedoms.

In the area of creating theoretical foundations for trusted AI, separate work is underway to develop a formal theory that defines the boundaries of applicability of AI methods and criteria for their reliability. The creation of a general theory will make it possible to move from an empirical search for solutions to the systemic design of secure algorithms and will create a basis for new fundamental research.

## 5. Important takeaways // Expert opinion

Security and trust are becoming core principles of modern AI, defining the boundaries of its application in critical areas, ranging from medicine and transportation to public administration. The focus is on aligning the goals of models with human values, resilience to attacks, data protection, and the development of transparent decision-making mechanisms. The modern AI safety paradigm goes beyond technical correctness and encompasses the entire development lifecycle: from dataset generation and training of models to their operation. The development of secure engineering practices (MLSecOps), model trustworthiness verification and testing, and explainability protocols provides the foundation for responsible implementation of AI that can act predictably and fairly.

The development of trusted and explainable AI has not only technological but also institutional significance.

International standards, protocols for audit and certification of AI systems are being actively developed to ensure transparency, sustainability, and accountability. The development of explainable models and tools for post factum analysis of decisions helps to strengthen public trust and reduce the risks of abuse and discrimination. In the long term, the "Safety, Trust and Explainability of AI" area is becoming a key element of the digital sovereignty architecture, contributing to the formation of a global culture of responsible AI use and the development of a secure, innovative economy.

The focus area is decisive for the mass implementation of AI

**27%** of the topics in this area are related to "aligning" AI goals with human ones, and developing correct value systems that provide the basis for further development and use of technologies.

FOCUS AREA 6

# Narrow AI

# FOCUS AREA 6

# Narrow AI

## 1. Overview of the focus area

Task-Specific AI (Narrow AI) is a class of AI systems designed to perform specific, well-defined functions. Unlike artificial general intelligence (AGI), which aims to mimic a broad range of human cognitive capabilities, narrow AI focuses on solving specific problems, which makes it possible to achieve high accuracy and practical utility in limited areas.

Its key advantage is that it already has practical use cases: from the industrial implementation of computer vision to language models in government services. Task-Specific AI has become a major driver of technological progress in AI due to its sustainable economic benefits and direct benefits to society and science. Machine learning is already widely used in a wide range of fields, from software development and planning in multi-agent systems to medical diagnostics, highlighting the significant impact of narrow AI solutions in industry and science.

The near future of AI does not appear monolithic: a shift is expected from the era of simple linear scaling of transformers to an era of architectural diversity. At the same time, transformers remain the foundation of many advanced systems: modern open LLMs (Qwen, LLaMa) use transformers, increasingly in the "mixture of experts" (MoE) form; in computer vision, foundation models such as Segment Anything Model 2 (SAM-2) and DINOv1-v3 are also based on transformer architectures; AlphaFold3 combines transformers with more complex specialized components.

At the same time, alternative approaches are also being actively explored, ranging from state-space models (SSMs), including hybrid Mamba + Transformer architectures (e.g., Nemotron, Granite), to neurosymbolic systems, world models, and graph neural networks.

In practical scenarios, highly specialized solutions are increasingly being used: from simple heuristic systems (for example, chatbots for university admissions campaigns) to more advanced tools with multilingual support and search integration (RAG). Researchers are exploring the potential of wearables to enhance human-machine interaction, including physiologically based stress monitoring and augmented reality (e.g., Meta Ray Ban Display glasses).

Research in task-specific AI is critical for both social sustainability and justice. The lack of reliability of the models, including a tendency to "hallucinations", limits their use in critical and high-risk areas (medicine, transportation, law). Improving trustworthiness and transparency will enhance public trust and ensure the secure integration of technologies in sensitive domains. In the social sector, narrow AI solutions are already automating the processing of citizen requests, reducing the workload on staff and preventing professional burnout. At the same time, the development of this area contributes to the democratization of access to AI and the expansion of the technology's reach by supporting languages with limited resources, thereby preventing the widening of social inequality.

Within this area, three sub-areas covering its basis were distinguished:

### 1. Computer vision (CV)

Computer vision technologies have reached industrial levels of maturity. A new area has emerged — video generation — where active progress began around 2023: commercial projects have emerged, video generation time has been reduced from hours to minutes, and quality continues to improve thanks to new architectures and optimization methods. The MagicTime model is better able to simulate physically based processes (such as plant growth) by training on time-lapse videos with detailed annotations (2000+ clips). One of the key areas in computer vision is the creation of foundation models trained on large corpora of data that are capable of solving computer vision problems on user data with a high level of quality. It is worth noting the progress in the field of segmentation and tracking of arbitrary objects, which has seen the emergence of models such as Segment Anything 2 and SAM, ob-

Special thanks for work on this focus area to:
D. Yudin, R. Baeza-Yates, E. Tutubalina

ject detection based on natural language queries, for instance, based on the YOLOE and Grounding DINO models, camera pose and depth map estimation from video using fundamental transformer models such as VGGT, answering questions (VQA, Visual Question Answering) and reasoning from images using visual-language models such as open-source Qwen-VL models, API-accessible models such as GPT-4o and GPT5, etc.

### 2. Natural language processing (NPL)

LLMs are being implemented everywhere, for example, in government services to automatically process citizen requests, reducing the workload on employees and preventing burnout. Multilingual training can improve the performance of LLM by exploiting common semantic structures across languages, but adding a large number of languages does not always result in a gain in quality. In practice, some companies, such as Cohere, have abandoned the 27 languages in favor of the six most important ones, which may be due to both reduced data collection costs and more stable final model metrics.

### 3. Other narrow AI technologies

Specialized AI methods are actively developing beyond CV and NLP. In software development, models are being used to optimize compilers; the concept of "AI for Science" is emerging to solve problems in pharmaceutics, physics, chemistry, biology, and materials science.

## 2. Overview of current developments

**Computer vision (CV).** Modern CV is characterized by the development of effective methods for analyzing visual information and the emergence of new generative capabilities. Models for generating images and videos from text (text-to-image, text-to-video), as well as methods for creating and editing 3D content, are actively developing.

**"Generative research and practical development in the field of AI *primarily involves text-to-image, text-to-video, text-to-visual models...*"**, as emphasized in the expert discussion, pointing to the key role of diffusion models in such tasks. Self-supervised learning methods are widely used in medical vision, allowing the use of unlabeled data, for instance, in radiology for image analysis. The latest advances in conformal forecasting and conformal risk management, integrating CVs into physician workflows (physicians participate in decision-making using AI models, and the physician remains informed). Practical CV systems have already been implemented in industry and medicine: automated sorting lines are in operation in production, and "second opinion" systems are used in diagnostics to assist doctors in analyzing images. New methods for extracting information from images are also being developed: for example, hyperspectral reconstruction allows for obtaining precise quantitative measurements from images, going beyond purely visual (aesthetic) tasks.

**Natural language processing (NLP).** There are strong advances in multi-task language models — GPT-5, Claude, Gemini (2023–2025) combine multimodality, improved reasoning, and few-shot learning. Large language models (LLMs) are being implemented in many areas. In particular, government services have begun using LLMs to process citizen requests, which reduces the workload on staff and helps prevent professional burnout. Highly specialized NLP models are finding application in the humanities: for example, AI is used to preserve cultural heritage through the recognition of ancient manuscripts and the study of language evolution based on historical texts.

**RAG (Retrieval-Augmented Generation).** Integrating search into the text generation process: the model queries external knowledge sources (e.g. documents) before generating a response. This approach has become fundamental for modern LLMs, as it allows for updating the knowledge of the model and referencing facts, reducing the number of "hallucinations". This approach is critically important in a rapidly changing world, because "if a model is trained on data from the previous year, naturally we won't get up-to-date data". An application example is "development of an AI agent for the RAGFlow admissions office", which successfully uses the RAG module to provide relevant answers in multiple languages. This highlights RAG's potential in areas where access to up-to-date data is critical, such as education, law, or public administration.

**Step-by-step reasoning (Chain-of-Thought, CoT).** Technique of step-by-step reasoning in the generation process. The models encourage to first explicitly write the chain of logical reasoning and then the final answer. This significantly improved LLM's ability to solve complex problems (mathematics, logic, everyday reasoning). The CoT effect is particularly evident in large models (over ~100 billion parameters) and is considered an emergent property — large models suddenly start

reasoning successfully if they are properly prompted with a chain of thoughts. However, there remain challenges related to hallucinations and the limitations of the models' inner "thinking". Development prospects include the integration of external tools and knowledge, as well as the development of systems capable of self-learning and self-diagnosis, simulating human "thinking, memory, planning". Integrating semantic networks or knowledge graphs into AI systems can improve the reliability of reasoning by providing access to structured facts and allowing filtering out inconsistent or false generations. This area is actively being researched as part of neurosymbolic AI and RAG systems.

**Structured output.** Generation of highly structured output (action plans, machine code, 3D molecular structures or tables) on request. For business applications, this is critical, as it allows LLM to be integrated into existing systems. However, achieving unconditionally correct formatting from a model is difficult: research shows that even advanced LLMs sometimes fail, especially on complex structures, requiring further improvements in the format's robustness. Developers implement special prompt templates and output control methods (such as constrained decoding) to improve the consistency of structured responses. Such models promote broad innovation and research. The integration of machine learning with structured rules, scientific knowledge, and ontologies (so-called neurosymbolic AI) is recognized as a promising approach. This makes it possible to constrain models and ensure more predictable and robust behavior, especially in areas where high accuracy and interpretability are required.

**Multi-agency.** Multi-agent systems and autonomous AI agents are being created for applied tasks: for example, highly specialized assistant bots for university admissions offices that speak several languages and can extract relevant facts from knowledge bases (an embodiment of the LLM + RAG combination). Moreover, LLM's agentic capabilities and Tool Calling capabilities allow language models to solve any application problems that can be solved using the given software tools. For instance, there are successful examples of LLM application in deep web information search (Deep Research) tasks, whereby calling a search engine, the language model synthesizes a report on a given topic. The flexibility of the Tool Calling mechanism allows LLM to be applied to a wide range of problems, which is an attractive direction for further research.

## 3. Research challenges that stimulate or limit the focus area development

### CHALLENGE 6.1

#### Limited and synthetic data

Highly specialized models of computer vision and natural language processing face a shortage of high-quality data in specialized areas. In medicine, it is difficult to collect samples for rare diagnoses, and in cybersecurity, it is difficult to collect reliable data on attacks or biometric anti-spoofing. Synthetic data generation is becoming critical, but its long-term implications remain unclear. Data deficiency limits the quality and reliability of models and reduces the possibility of their implementation in practice. If an effective way to replace real data with synthetic data is found, it may pave the way for scalability and accelerate the adoption of AI in complex domains. In this regard, methods for generating synthetic datasets are being actively developed, including those based on specialized simulators, and consortia are being formed to exchange rare medical and industrial data. In parallel, approaches to validating the quality of synthetics are being explored to minimize the risks of distorting training results.

### CHALLENGE 6.2

#### Creating efficient spatial representations for solving computer vision problems

The key problem is still the lack of universal methods for representing spatial data from various sensors for computer vision tasks. Despite the development of multimodal models and three-dimensional representations (NeRF, Gaussian Splatting), existing approaches do not ensure effective integration of semantic features and remain unsuitable for tasks requiring high response speed. There remains a need to develop adaptive methods that can combine information from disparate sensors and support complex spatial reasoning, which is necessary for practical application in real-world conditions.

### CHALLENGE 6.3

#### Hardware-software joint optimization of computer vision systems

High-precision operation of computer vision systems (in measuring equipment, medical imaging) requires

not only the development of algorithms, but also the improvement of sensors, optics, and equipment. Without comprehensive optimization, it is impossible to achieve mass availability of such systems. Co-optimization of hardware and algorithms can transform expensive prototypes into widely applicable solutions, accelerating the adoption of CV in medicine and industry.

This area is critical for AI to move beyond the laboratory. Interdisciplinary projects are being created that bring together engineers and AI specialists. New sensors and specialized chips adapted to specific computer vision tasks are being developed.

### ϟ CHALLENGE 6.4

**Languages with limited resources, including rare languages**

Most of the world's languages are poorly represented in digital corpora, which makes them under-represented in large language models. This creates a digital divide and reduces the versatility of NLP technologies. Support for small (rare) languages is critical for the global availability of AI. Solving this problem will significantly expand the reach of technology and reduce social inequality, opening up opportunities for new educational and cultural services. To respond to this challenge, projects are being created for collecting and marking up texts for languages with limited resources. Methods for synthetic corpus replenishment and multilingual learning, including knowledge transfer from more common languages, are explored.

### ϟ CHALLENGE 6.5

**Using alternative architectures**

Despite the dominance of neural network transformers, researchers are considering unconventional approaches for narrow CV and NLP tasks. For example, neuromorphic spiking networks and optical neural networks are being studied. While these solutions are still at a fundamental stage and not mature enough to compete with mainstream methods, they open up new directions for the development of AI architectures. The development of alternative architectures can overcome the limitations of current models, improve energy efficiency, and open up new areas of application.

### ϟ CHALLENGE 6.6

**Bridging the gap between simulation and the real world (Sim2Real) for solving complex problems in robotics and autonomous vehicles**

In robotics, there remains a gap between simulation and the real world (sim2real). Algorithms that work well in simulators are exposed to the real world without knowledge of the true physical properties of objects. Overcoming this barrier requires methods that are not only robust to environmental variability and sensor noise, but also are able to perceive and understand physical properties. Existing approaches integrate physical knowledge of the world through learning on datasets in a question — answer format and have limited transferability of this knowledge directly to the objects of manipulation. The problem is exacerbated by a lack of high-quality real-world data for training robots and limitations in the computing power of autonomous platforms. Ensuring reliable and stable operation of robots in real-world conditions remains an unsolved global challenge. Integrating reliable information about physical properties into training data is a crucial step to ensure that the robot can follow instructions predictably, safely, and effectively.

## 4. Long-term research tasks

### ★ TASK 6.1

**Development of computer vision methods for simulating real-world scenarios (including embodied AI)**

Particular attention is paid to creating computer vision methods that are robust to complex real-world scenarios and capable of bridging the gap between simulation and reality (Sim2Real). For this purpose, methods of domain randomization, imitation and active learning, as well as Vision-Language-Action (VLA) architectures are being worked out. Such methods enable models to learn from their own actions in a simulated environment and transfer this knowledge into the physical world. Such systems will be critical for robotics, autonomous vehicles, and industrial applications that require not only high perception accuracy but also guaranteed safety of behavior.

### ★ TASK 6.2

**Creation of foundation VLMs for various CV tasks (such as SAM, Dyno v2) (solving problems of classification, detection, segmentation with an open list of classes)**

Such models must possess multimodal spatial representations that integrate information from images,

videos, texts, and sensor data (including lidars and IMUs). Development is progressing towards contrastive and masked learning on multimodal data, the fusion of 2D and 3D representations, and the integration of physical plausibility mechanisms into the training process. It is expected that such systems will be able to understand scene context, work with previously unseen objects, and perform spatial reasoning, representing a step towards creating universal perception systems for robotics and embodied AI.

### ★ TASK 6.3

**Research and development of efficient training and inference methods for natural language processing architectures (including AutoML)**

This research focuses on optimizing architectures using AutoML, parameter-efficient approaches (LoRA, adapters), knowledge distillation, and inference acceleration techniques, i.e. quantization, pruning, and mixed precision. A key priority is reducing energy consumption and adapting models for distributed and specialized computing environments. Solving this task will enable will make it possible to create more affordable, faster and updatable language systems applicable to a wide range of tasks, from intelligent assistants to specialized industry AI applications.

### ★ TASK 6.4

**Research and development of efficient training and execution methods for RecSys, S2T and TSA architectures (including automatic learning)**

Finally, the unification of approaches to training and executing models in various fields, such as recommender systems, speech recognition and time series analysis, remains an important area. Current research focuses on developing architectures that ensure efficient processing of streaming data and low latency with high prediction quality. AI specialists use streaming transformers, state-space models, and contrastive self-learning methods, as well as parameter-efficient tuning, which allows large models to be adapted to narrow domains and programming languages without complete retraining. The outcome will be the creation of universal architectures and algorithms that guarantee adaptability, energy efficiency, and accuracy under limited resources, thereby enhancing the scalability and accessibility of modern AI solutions across different sectors of the economy.

## 5. Important takeaways // Expert opinion

The scientific and technical community exhibits a wide spectrum of views on the priorities and development paths for Narrow AI — ranging from fundamental theoretical research to practical implementations. The key positions and disagreements voiced by experts are outlined below.

Researchers identify two key development paths for Narrow AI:

1) unification, which relies on foundation language models and relatively limited fine-tuning;

2) specialization, which involves the development of domain-specific models.

The second path often allows for greater reliability, while the first is more scalable and easier to maintain. Consequently, it is logical to use the specialized approach in fields with a high cost of error (e.g., medical diagnosis, cybersecurity), and the unified approach for less critical applications (e.g., chatbots, general text processing).

One of the key problems limiting the applicability of the LLM-based approach is reasoning unfaithfulness. Despite their superficial plausibility, the models' textual reasoning often incompletely or inaccurately reflects the actual decision-making process. They may omit key factors that influenced the decision, tailor explanations to fit a plausible but incorrect answer, and mislead about the crucial factors behind the response. Therefore, research aimed at enhancing the faithfulness of language model reasoning is highly promising. Success in this area will gradually enable the expansion of their use into domains with a higher cost of error, particularly when combined with domain-specific and more transparent approaches.

Focus on applied research. Many experts emphasize the importance of applied work on AI. Particular attention is paid to the holistic optimization of "sensors-neural networks-processors" for specific tasks. This science-informed ML approach aims to bridge the gap between laboratory prototypes and industry-ready solutions. In practice, this means that success is seen in the tight integration of fundamental research with industry needs, ensuring that new models immediately account for the characteristics of real-economy data and hardware.

**Foundation models and the trend toward efficient spatial reasoning.** Modern foundation models (particularly VLMs), whose development has seen significant progress, still perform poorly on tasks involving reasoning, question-answering, and object localization in 3D space. The development of such models and architectures capable of high-quality spatial reasoning is a current trend that could lead to the creation of advanced environmental perception systems for robots, autonomous vehicles, virtual/augmented reality systems, smartphone applications, and similar domains.

**Efficiency and accuracy.** The need to develop more efficient training and inference methods is separately emphasized. In particular, some experts view computer vision tasks as a distinct area for optimization: it is necessary to achieve acceptable speed and energy efficiency of CV models without a significant loss of accuracy. This balance is critical for deploying CV algorithms in resource-constrained devices and for real-time video processing.

**Multilingual barrier.** The problem of supporting low-resource languages is recognized as a serious challenge by the entire community. Experts deem it necessary to adapt existing LLM architectures and generate new corpus data to bring as many of the world's languages as possible into the AI sphere. A key direction is the development of higher-quality translation models — a task that remains unsolved. Such models can be used both for generating synthetic multilingual datasets and for enhancing cross-cultural communication, thereby expanding the accessibility and effectiveness of NLP technologies. Without these efforts, NLP technologies risk exacerbating inequality by leaving the majority of language communities without modern tools.

**Beyond neural networks.** Some researchers are calling for a move beyond the currently dominant paradigm of deep neural networks. There are appeals for a broader approach and the exploration of alternative AI models (symbolic, evolutionary, etc.) to ensure valuable methods outside the neural network mainstream are not overlooked.

**Perspective on Narrow AI integration.** Some experts think that future progress in specialized AI may be linked to breakthroughs in foundation models. *"I believe that with a genuine breakthrough in natural language processing models, all tasks in area 6.3 could be solved. However, for this, we need precisely a breakthrough in the core NLP model",* says Shunquan Tan. He emphasizes that a qualitative improvement in core language models could automatically propel the development of related Narrow technologies (speech recognition, recommender systems, etc.). This opinion highlights a key disagreement within the community: some see the future of Narrow AI in specialized models for each specific task, while others see it in the universalization of foundation models with their subsequent adaptation for specific applications.

---

The most application-oriented focus area, encompassing tasks critical for specific applications and industries

---

The most dynamically changing area in terms of the research landscape is the number of tasks themselves, which is constantly growing, and their prospects are changing significantly in the context of volatile external circumstances

---

**62%** of tasks are related to computer vision technologies

---

The current landscape is dominated by the deep neural network paradigm. However, this field requires a broader approach, incorporating alternative AI models (such as symbolic, evolutionary, and others)

---

FOCUS AREA 7

# Control, decision-making, and agentic/ multi-agent systems

# FOCUS AREA 7

# Control, decision-making, and agentic/ multi-agent systems

## 1. Overview of the focus area

The field of AI for control, decision-making, and agentic/ multi-agent systems has undergone rapid evolution: from rule-based programs to complex autonomous systems that are self-learning and capable of perception, reasoning, and action in complex environments. This transformation has been largely driven by progress in deep learning, reinforcement learning (Reinforcement Learning, RL), and multi-agent coordination. The core of this field lies in developing AI systems capable of autonomous decision-making, control, and interaction with other agents or the environment to achieve specific goals, often by optimizing long-term rewards. The boundaries of the focus area are constantly expanding, encompassing areas from cooperative robotics and swarm intelligence to game AI and complex decision-making across various sectors.

The main subareas of this area are as follows:

1. Reinforcement learning

This approach involves an agent learning optimal strategies by maximizing cumulative reward through interaction with its environment. It is a fundamental paradigm for building AI systems for control and decision-making.

2. Agentic systems

These are holistic, autonomous AI entities capable of perception, learning, and adaptive action. Currently, through the design and use of complex architecture, agents solve a wide range of applied problems.

3. Multi-agent systems

They facilitate the interaction, coordination, and cooperation of multiple agents to solve complex tasks. This encompasses methods for communication, competition, and collective decision-making.

Historically, AI for control and decision-making began with classical control theory and expert systems based on predefined rules and programming. The advent of machine learning, particularly deep learning and reinforcement learning, marked a significant shift towards data-driven and adaptive approaches. This also introduced trial-and-error learning using Markov Decision Processes (MDPs) and Markov Games (MGs) to model agent-environment and multi-agent interactions. In recent years, the focus has shifted to scalability in large state-action spaces, ensuring safety in real-world applications, and developing generalized learning capabilities. A clear trend is emerging towards creating AI agents capable of self-learning and operating in diverse, open, and dynamic environments. Such agents are expected to orchestrate external services and prioritize task resolution. The integration of Large Language Models (LLMs) and Foundation Models (FMs) into control tasks has given rise to a whole range of new research directions, the ultimate goal of which is to endow agents with enhanced abilities for reasoning, planning, and communication.

*"In recent years, a trend has emerged towards developing intelligent agents for robotics and autonomous vehicles. The emergence and development of foundational VLA (Vision-Language-Action) models, such as π0 [a], Gemini Robotics [b], and Gr00t [c], are gradually paving the way for general-purpose robots. These could become universal domestic assistants and help automate routine operations in manufacturing, among other applications".* — Dmitry Yudin, Principal research scientist, Cognitive AI Systems Lab, AIRI Institute.

The following key milestones have formed the foundation for the development of this field:

- DeepMind's AlphaGo breakthrough. It demonstrated the power of RL in complex strategic environments and inspired further research in the field of general game AI and control systems.
- Development of Large Language and Foundation Models. The rapid development and widespread availability of LLM and FM have radically transformed agent-based and multi-agent systems, providing powerful capabilities for endowing models with reasoning, planning, and natural language communication capabilities.

Special thanks for work on this focus area to:
E. Burnaev, A. Darwish, D. Yudin

- Ongoing breakthroughs in robotics: Robots and autonomous vehicles are mastering increasingly complex manipulation, dexterous locomotion, and operation in unstructured environments. This brings AI out of simulations and into the physical world, with direct economic consequences.
- Increased focus on multi-agent cooperation and federated learning. The growing need for distributed intelligence and privacy-preserving AI has stimulated active research in Federated Multi-Agent Reinforcement Learning (FMARL), particularly in sectors like healthcare and finance.
- Development of generative AI agents. The creation of frameworks and technologies for forming and training LLM-based agents has accelerated the widespread adoption of AI across various industries.
- Development of intelligent agents capable of automating hypothesis generation, experiment execution, code writing, and scientific paper authoring (e.g., Google DeepMind's AlphaEvolve, Sakana AI's AI Scientist). Such solutions have the potential to accelerate scientific discovery in mathematics, computer science, engineering, and more.

AI for control, decision-making, and agentic systems is one of the most critical focus areas, as its development is fundamental to the next generation of autonomous technologies poised to transform nearly every sector of industry, the economy, and daily life.

- Impact on industry, transportation, and safety. Driven by progress in robotics and autonomous systems, this includes the accelerated adoption of self-driving vehicles, UAV swarms, and advanced robotics. This will lead to a radical transformation of transportation, manufacturing, and operations in hazardous environments.
- Impact on quality of life. Through the transformation of personalized services, digital ecosystems, and cybersecurity, AI agent technologies are structurally changing digital landscapes. This enables advanced game AI, intelligent interfaces, and personalized service systems — from adaptive educational platforms to clinical assistants. In cybersecurity, they can contribute to more proactive threat detection and neutralization, helping to protect critical digital assets.
- Impact on the economy as a driver for economic growth by boosting productivity, reducing operational costs, and creating new industries and services. It enables a high degree of automation in manufacturing, logistics, and infrastructure management. The economic effects are expected to be large-scale, impacting global markets and creating new competitive dynamics.
- Impact on the labor market. The widespread adoption of AI agents and autonomous systems is predicted to transform labor markets by automating routine tasks and creating demand for new skills in AI development, maintenance, and governance. While some jobs may be displaced, new opportunities are expected to emerge, necessitating large-scale retraining and upskilling programs.

## 2. Overview of current developments

The current state of AI for control, decision-making, and agentic/multi-agent systems is characterized by rapid progress in Deep Reinforcement Learning (DRL) and the growing integration of Large Language Models (LLMs). Deep Reinforcement Learning has demonstrated high efficiency in decision-making for complex environments — from games and robot control to resource management and healthcare. In the last 1-2 years, a marked shift has occurred towards developing unified AI agents that integrate self-learning, planning, and interaction into a single architecture. Priority research focus areas include enhancing the robustness, safety, and generalization capabilities of reinforcement learning agents. Specifically, there is active development in Risk-Constrained RL, which focuses on building behavior strategies that account for safety constraints and minimize failures. This is critically important for safety-sensitive applications like autonomous driving.

Furthermore, the "Agent AI" concept is gaining significant traction. These are systems that integrate large foundation models into an agent's behavioral loop, leading to the formation of a more holistic intelligence. Multi-task learning methodologies are being used to create adaptive and versatile agents for robotics, game AI, and healthcare. In multi-agent systems, special attention is paid to coordination, communication, and decentralized decision-making, fueled by advances in LLM-oriented Multi-Agent Reinforcement Learning (MARL) and federated learning. The development of multi-agent architectures based on LLM-driven agent frameworks, including Actor-Critic and role-based models, is a major trend and paves the way for scalable solutions to problems requiring cooperation and competition.

## 3. Research challenges shaping and limiting the focus area development

### CHALLENGE 7.1

#### Scalability and complexity in multi-agent systems

A key problem in Multi-Agent Reinforcement Learning (MARL) is dealing with large state and action spaces, whose dimensionality grows exponentially with the number of agents and their individual complexity. This issue, known as the "curse of dimensionality", significantly reduces the effectiveness of existing approaches as environments become more complex. At the same time, research into multi-agent architectures extends beyond an engineering challenge — it touches upon fundamental questions about the nature of collective intelligence. Phenomena such as phase transitions and self-organization in multi-agent systems could hold the key to understanding how qualitatively new forms of intelligent behavior emerge from the interaction of relatively simple agents.

Solving this challenge would lead to breakthroughs in swarm intelligence, cooperative robotics, and large-scale resource optimization.

In the context of this problem, researchers are actively exploring the possibilities of Hierarchical Reinforcement Learning (HRL) for decomposing complex problems into solvable subproblems. AI agents are increasingly integrating Retrieval-Augmented Generation (RAG) modules and iteratively using LLMs to refine models through code and prompts. Agents are being designed to be more proactive, capable of orchestrating external services and prioritizing tasks autonomously. New models, such as MapGPT, are being developed for decentralized multi-agent planning. A significant trend is self-play, where AI models generate vast amounts of data by interacting with each other. This leads to performance improvements that can surpass training on human-generated data.

### CHALLENGE 7.2

#### Safety, reliability, and trust in AI agent operations

Despite the empirical successes of RL, transferring these results to real-world applications remains difficult due to the challenges of ensuring safety, robustness, and trust.

This is especially critical in high-stakes domains like autonomous driving or industrial control systems. Establishing clear regulatory frameworks, boundaries of liability, oversight mechanisms, and methods for auditing AI decisions is paramount for the responsible deployment and public acceptance of increasingly autonomous AI systems.

Public reaction to AI failures is asymmetric: an autonomous system may be statistically safer than a human, but a single failure can undermine trust in the entire technology. Therefore, solving this challenge would unlock significant opportunities for real-world AI applications in transportation, healthcare, and critical infrastructure.

Significant research efforts are focused on risk-constrained reinforcement learning and formal verification of safety strategies, the use of distributed reinforcement learning to more fully capture uncertainty and risk, and cost-constrained safe reinforcement learning (Safe-RL) to ensure the safety of agents through bi-criterial optimization of the specified criteria.

### CHALLENGE 7.3

#### Simulation-to-Reality gap for Embodied Agents

Transferring learning from virtual (simulated) to real environments remains a significant obstacle for embodied AI agents, particularly in robotics. Behavior policies learned in high-fidelity simulations often underperform when deployed in the physical world due to discrepancies in simulated vs. real physics, sensor data noise, and environmental unpredictability. This sim-to-real gap makes training robots directly in real-world settings difficult and resource-intensive.

This problem significantly slows the development and deployment of intelligent robots and physical autonomous systems. It requires extensive and often risky real-world data collection and subsequent fine-tuning, limiting the scalability of robotics research and practical applications. Overcoming it would radically transform robotics, enabling faster development cycles and expanding applications in manufacturing, research, and service industries. For AI in general, it stimulates research in domain adaptation and the development of robust methods for learning from limited real-world interaction.

Methods such as domain randomization, where simulation parameters are varied widely, are used to train more robust policies that generalize better to reality. Meta-learning and transfer learning are being studied to enable agents to adapt quickly to new real-world conditions

with a minimal amount of real experience. Research into universal physical action models and multimodal Vision-Language-Action (VLA) models aims to create more generalized representations that are less sensitive to the discrepancies between simulation and reality.

## 4. Long-term research tasks

### ★ TASK 7.1

#### Developing Universal Multimodal Models integrating text and other modalities: Vision-Language-Action (VLA)

Multimodal systems, particularly in robotics, are poised to become one of the leading research directions over the next decade. The objective is to create truly universal multimodal agent models capable of seamlessly integrating perception (e.g., vision), natural language understanding and generation, and physical actions (Vision-Language-Action, or VLA models). Such models must demonstrate multitasking capabilities, self-learning, and efficient operation in open, unstructured environments without the need for task-specific fine-tuning.

Solving this task requires large-scale pre-training on heterogeneous modalities. Key methodological approaches include self-supervised learning, masked autoencoders, and contrastive learning applied to integrated streams of visual, language, and behavioral (action) data. Reinforcement learning will be critically important for enabling self-learning and adaptation in dynamic environments, with Hierarchical RL potentially providing a structure for managing complex behavioral scenarios. Furthermore, this task faces a significant data scarcity challenge, making the generation of synthetic data and the use of self-play promising solutions to move beyond the limitations of human-curated datasets.

Successfully addressing this task will pave the way for highly versatile robots for domestic, industrial, and research applications, as well as intelligent personal assistants capable of more "human-like" physical and verbal interactions.

### ★ TASK 7.2

#### Developing effective methods for agent knowledge acquisition through environmental interaction

This task focuses on creating algorithms and frameworks for Multi-Agent Reinforcement Learning (MARL) that provide formal guarantees of safety, robustness, and ethical compliance, particularly for real-world and safety-critical applications. The goal is to move beyond purely empirical effectiveness, ensuring multi-agent systems can operate within specified risk boundaries, avoid catastrophic failures, and exhibit predictable behavior even in complex, uncertain, and hostile scenarios.

Successful resolution will yield a deep understanding of how to formally guarantee desired properties in complex adaptive systems. This in turn will ensure the widespread implementation of autonomous multimodal systems in highly sensitive fields such as autonomous transportation systems, smart energy grid management, and medical robotics, where failure is not an option.

### ★ TASK 7.3

#### Research and development of multi-agent systems and investigation of phase transition phenomena in multi-agent systems

The objective is to design multi-agent architectures that foster the emergence of advanced collective intelligence and coordinated behavior through hierarchical organization and decentralized learning, often drawing inspiration from biological systems and complex adaptive systems theory. This involves developing "master systems" capable of dynamically creating and managing specialized, narrowly-focused agents, as well as studying phase transition phenomena that occur as the number of agents increases.

Hierarchical Reinforcement Learning (HRL) allows agents to learn at different levels of abstraction and time scales. Genetic algorithms and evolutionary methods can be used for evolutionary selection and optimization of agent roles and communication protocols within a hierarchy. Modern communication protocols and emergent languages in multi-agent systems enable enhanced coordination and information sharing. Game theory and mechanism design can be used to incentivize cooperative behavior and manage competition.

Solving this task will enable the creation of scalable, flexible, and robust multi-agent systems capable of adapting to radical environmental changes and solving problems far beyond the capabilities of individual agents. This will lead to advanced UAV swarms for complex surveillance and search-and-rescue operations, intelligent traffic management systems with real-time adaptation, and highly distributed robotic complexes for large-scale construction or environmental monitoring.

## 5. Important takeaways // Expert opinion

The development of AI for control, decision-making, and agentic and multi-agent systems is characterized by a significant shift towards integrated, autonomous, and adaptive intelligence. The rapid evolution of deep reinforcement learning (DRL) and the integration of foundation models, particularly LLMs, are key trends. This is leading to the creation of agents with advanced perception, reasoning, and the ability to work with various modalities in complex environments.

The focus has shifted from specialized, rule-based systems to universal, trainable architectures capable of operating under environmental uncertainty and dynamics.

The increasing complexity and scale of real-world problems necessitate multi-agent solutions, stimulating research in robust coordination, communication, and decentralized decision-making. Overall, the development pathway is aimed at creating more general, intelligent, and autonomous AI entities capable of efficient learning, adaptation, and interaction, thereby bridging the gap between theoretical advances and practical deployment in the real world.

Furthermore, addressing security challenges in multi-agent systems, particularly concerning collusion and coordinated attacks, is becoming increasingly critical as agents interact across various internet platforms.

Technological progress alone is not enough; addressing issues of accountability and regulatory frameworks is becoming a critical factor as well. The success of this area will be determined by the research community's ability to bridge the gap between impressive performance in simulations and reliable operation in reality.

Strong emphasis on practical, efficient, and reliable solutions for autonomous systems

Key development trends in this area are associated with a transition towards integrated, autonomous, and adaptive intelligence

**40%** of the area foundation is research related to reinforcement learning

Agentic and multi-agent systems are rapidly evolving, driven by growing societal demand for automation and autonomy. For entire classes of tasks, agents could become a sort of new evolutionary form of "conventional AI"

A major challenge for this is given the demand for automation and autonomy, is ensuring the robustness and reliability of the developed technologies and systems

FOCUS AREA 8

# Elements of AGI

# FOCUS AREA 8

# Elements of AGI

## 1. Overview of the focus area

"Elements of AGI" is a field of research where AI is evolving from narrow tasks towards systems capable of reasoning, lifelong learning, integrating knowledge across different forms, and operating in complex environments. The practical focus lies on fostering robust properties in future systems: credible reasoning and self-verification, continuous learning through environmental interaction, hybrid neuro-symbolic approaches, embodiment, multi-agent capabilities, and neuro-inspired architectures. Based on foresight studies and expert interviews, a consensus emerges: the definition of AGI remains fluid, and progress is driven not by a single "leap" but by the convergence of the aforementioned research lines. In the next 3–5 years, the most significant momentum is anticipated in agentic/multi-agent systems and continuous learning; in the longer term, progress is expected in neuro-symbolic AI and brain-inspired models.

The core of this area is the transition from "models as tools" to systems demonstrating adaptive versatility: the ability to reason and plan reliably, learn from experience, and transfer knowledge to new domains. These properties are not reducible to a single algorithm; they emerge from the interaction of reasoning/reflection, memory and personalized context, environmental learning, hybrid knowledge representations, and agentic action.

The boundaries of this field are defined by the shift away from narrow optimization for a fixed dataset and towards robust performance in an open world: operating "out-of-distribution", explainability and verifiability of inferences, the capacity for long-term memory, and accurate self-assessment of results. The practical criterion for success is the systems' ability to solve complex applied tasks (in science, medicine, engineering design) under resource and data constraints while maintaining safety and trust.

The following key research subareas are currently distinguished:

### 1. Reasoning and reflection

Focus on enhancing the quality of reasoning (including multimodal), causal understanding, planning, and self-verification mechanisms — reflection, calibrating confidence, and verifying chains of inference. This also includes practices for explainability and certifiable robustness of inferences.

### 2. Continuous learning

Transition from static training to continuous learning: active and curriculum learning, online/offline RL, self-play, and data acquisition from the environment. The goal is to overcome the limitations of fixed datasets and achieve adaptation to changing conditions without catastrophic forgetting.

### 3. Hybrid AI

Integration of deep learning with ontologies, knowledge graphs, and logic for interpretability and robustness. The neuro-symbolic approach is combined with tool usage and retrieval-augmented practices, which enhances the accuracy and controllability of inference.

### 4. Embodiment and multi-agent systems

Learning through action in real/virtual environments, coordination of multiple agents, and communication between them. Foundation models (LLMs/VLMs) serve as the base layer, providing perception and planning; the key challenge is the simulation-to-reality transfer (sim2real) and behavioral safety.

### 5. Brain and mind simulation

Neuro-inspired models — from spiking neural networks to the simulation of brain circuits — are considered a long-term source of energy efficiency, noise resilience, and new principles for memory/generalization. In parallel, AI is being used to interpret neural data (EEG, etc.), although such data is scarce and noisy.

Historically, the area has evolved from symbolic AI (expert systems, ontologies) through statistical machine learning to deep learning and transformers — this is the trajectory that gave rise to the hope for general models. However, scaling revealed limitations (hallucinations, weak reflection, high costs), which renewed interest in hybrid approaches, continuous learning, and agency. In games, self-play (e.g., in Go) demonstrated the power of learning "from the environment", while early neuro-inspired ideas showed the potential of biological principles for

Special thanks for work on this focus area to:
I. Oseledets , Y. Tian, S. Kolyubin

efficiency and long-term memory.

The most significant influences on the area development over the past 5 years have been:

- **ChatGPT and subsequent LLM-based agents.** A moment of mass recognition for the potential of foundation models: dialogue, tool use, and chains of reasoning. This sharply raised requirements for AI reliability, explainability (XAI), and risk management, creating a demand for elements of AGI.
- **RAG/Tool-use as a hybrid standard.** The widespread adoption of retrieval approaches and integration with external tools/knowledge bases established neuro-symbolic practices in the industry. This improved. factual accuracy and output controllability, confirming the importance of hybrid AI.
- **Chain-of-Thought and reflection practices.** The proliferation of techniques for explicit reasoning, self-verification, and confidence calibration set a new standard for reasoning tasks. In response, research into certifiable robustness and reflection intensified.
- **Embodied/Multi-agent systems and learning worlds.** The growth of platforms and approaches where agents learn to act and coordinate in dynamic environments demonstrated the practicality of shifting from static datasets to experience. The sim2real problem and behavioral safety became central agenda items.
- **Spiking and brain-inspired neural networks as a long-term vector.** A resurgence of interest in energy-efficient and biologically-motivated architectures, as well as using AI to read/interpret neural data. Despite data noise and scarcity, this direction has solidified as promising.

## 2. Overview of current developments

In the Reasoning and Reflection subarea, the focus is shifting towards methods for enhancing inference reliability: structured reasoning chains, self-assessment, fact verification, XAI approaches, and practices for certifiable robustness.

In the Continuous learning subarea, online learning and self-play are gaining prominence, with active exploration of overcoming data saturation through experience acquisition and the generation of high-quality synthetic data.

In the Hybrid AI subarea, RAG/Tool-use practices and integration with ontologies/knowledge graphs have become a de facto standard for tasks where accuracy and auditability are crucial.

In the Embodiment and Multi-Agent Systems subarea, there is a growth in platforms for learning through action and agent coordination; the key barriers remain safety and real-world transfer.

In the Brain and Mind Simulation subarea, the search for energy-efficient architectures and methods for interpreting neural data continues; the direction is developing, albeit constrained by data availability.

Over the last 1–2 years, work on reasoning and reflection has intensified; RAG and tool-using agents have become established in products; interest in multi-agent systems and learning in environments with feedback is growing actively. In parallel, an efficiency agenda is taking shape: there is a growing need to reduce the cost of training and inference, and to improve memory management and personalization.

## 3. Research challenges shaping and limiting the focus area development

### CHALLENGE 8.1

#### Reasoning and reflection

Core Challenges: verifiable inference chains, confidence calibration, robustness to distributional shifts. Scale of Importance: critical for medicine, law, science; the solution lies in XAI (Explainable AI), validation procedures, reflexive cycles, and certifiable methods. Required Measures: development of evaluation and auditing standards, benchmarks, and industrial-grade validation practices.

### CHALLENGE 8.2

#### Data saturation and continuous learning

Modern AI systems are transitioning from a static "train-then-deploy" paradigm to a model of continuous data and experience accumulation. This makes it possible to overcome the saturation of pre-existing datasets and avoid knowledge degradation during fine-tuning. The key objective is to teach models to extract information from their environment,

enrich their own experience, and retain previously acquired skills. This approach forms the basis for creating robust, long-term AI agents capable of learning throughout their entire lifecycle. To achieve this, methods such as online reinforcement learning (online RL), self-play, active and curriculum learning are employed, alongside the generation of high-quality synthetic datasets that enable continuous knowledge updating.

### ⚡ CHALLENGE 8.3

#### Sim-to-real and safety of embodied/multi-agent systems

A central challenge is the transfer of skills acquired in simulators to the real world. For autonomous and robotic systems, it is critical to guarantee the safety, predictability, and reproducibility of agent behavior outside laboratory conditions. Research is focused on developing methods for domain randomization, creating high-fidelity simulators, testbeds, and certification protocols that minimize the risks of incorrect system behavior when interacting with the physical environment and humans.

### ⚡ CHALLENGE 8.4

#### Deficiency and protection of neurodata

The advancement of neuro-inspired architectures and brain-reading technologies is hampered by the scarcity, noisiness, and sensitivity of neurodata. These data are difficult to acquire and anonymize, which slows down progress in creating energy-efficient and cognitively-motivated models. Solving this challenge requires the development of standards for safe collection and anonymization, as well as the creation of data compression methods and compact signal representations. This will enable the use of neurodata without violating ethical and legal norms.

### ⚡ CHALLENGE 8.5

#### Computational efficiency and scalability

The increasing size of models and the rising cost of their training have made energy efficiency and the optimal use of computational resources a critical issue. The goal is to reduce the costs of training and inference without sacrificing quality, thereby ensuring the scalability of technologies and the accessibility of AGI components. This direction involves the optimization of computational processes, the development of new types of hardware and architectures (including distributed and energy-efficient solutions), as well as the creation of methods for dynamic resource allocation in complex computing systems.

## 4. Long-term research tasks

The advancement of AI systems towards greater autonomy and complexity generates new fundamental challenges in the areas of safety, reliability, trust, and explainability. These challenges require deep research that goes beyond existing approaches, touching upon both the theoretical foundations of machine learning and the applied aspects of AI's interaction with humans and the environment. Below are key research tasks identified through expert discussions.

### ★ TASK 8.1

#### Enhancing AI model generalization and adaptability through continuous learning

The core challenge lies in overcoming a fundamental limitation of modern AI models: their inability to effectively generalize to new, previously unseen domains or tasks without complete and costly retraining. As an expert notes, *"if we train a neural network for one domain, there is no guarantee it will perform equally well in other domains".* It is necessary to develop mechanisms for "elasticity in forgetting and remembering knowledge" that would allow models to continuously adapt to new tasks while retaining previously acquired experience, a concept known as "learning without forgetting".

To achieve this, research is focused on continual learning methods, new loss functions that account for the preservation of prior knowledge, domain adaptation, and cognitively-inspired approaches based on the principles of human memory and forgetting. The implementation of such solutions will enable the creation of general, robust, and adaptive AI models capable of learning from small data and continuously improving in dynamic environments, which is particularly critical for robotics and personalized systems.

### ★ TASK 8.2

#### Reasoning: quantitative uncertainty estimation and reflection

Modern large language models, despite their ability to answer complex questions, often make "very, very stupid mistakes" and are prone to "hallucinations" — generating factually incorrect information presented with high confidence.

The challenge is to develop methods that enable AI models to adequately assess, quantify, and verbalize their level of confidence or uncertainty in their answers. The goal is to teach models to "refuse to answer" or signal a lack of confidence, which is critically important for their safe deployment.

A variety of approaches are being explored. These include unsupervised methods that leverage the model's internal statistics (e.g., entropy, token probabilities) to estimate confidence, and supervised methods that train auxiliary models to identify errors. Furthermore, research involves "black-box" techniques, which analyze the consistency of answers without accessing internal states, and "white-box" methods, which directly utilize the model's internal parameters. This will enable the creation of more reliable and self-critical AI systems capable of interacting with humans safely, "rejecting" dubious queries, and enhancing the faithfulness of reasoning in critical areas.

### ★ TASK 8.3

#### Development of multi-agent systems and optimal governance for complex tasks

This task is focused on creating complex AI systems composed of multiple interacting agents capable of coordination, cooperation, and adaptation in dynamic environments. The scope extends beyond conventional Reinforcement Learning (RL) and includes the development of optimal governance methods for solving problems in sampling, generation, and fine-tuning of large models. It is necessary to establish the theoretical and practical foundations for governing the behavior of such systems, including ensuring their predictability and safety.

Solution methodologies involve the advancement of reinforcement learning and optimal governance algorithms.

This will enable the creation of autonomous multi-agent systems capable of solving complex problems, accelerating scientific discovery, boosting productivity, and providing more precise governance and fine-tuning of large models.

## 5. Important takeaways // Expert opinion

The evolution of the area is defined by a fundamental shift from building narrow, specialized models towards developing more general, adaptive, and autonomous systems. A key trend is the convergence of several research lines: robust reasoning, continual learning, hybrid approaches, and multi-agent systems. We are witnessing a transition from static training on fixed datasets to continuous learning through environmental interaction. This shift is driven by the saturation of existing data and the necessity for models to adapt to dynamic real-world conditions. Hybrid neuro-symbolic approaches, such as RAG and the use of external tools, are gaining significant importance, as they enhance the accuracy and controllability of AI inferences.

There is a growing recognition that scaling existing architectures has exposed their fundamental limitations, including a trend toward "hallucinations" and weak self-reflection. This has made quantitative uncertainty estimation one of the most central and active areas of research. In response, investigations into self-checking and confidence calibration have intensified. That is a necessary step toward building reliable and safe AGI systems. Furthermore, the focus area of multi-agent systems and embodied AI is advancing rapidly. Here, foundation models serve as a base layer for perception and planning, with the core challenge becoming the transfer of skills from simulation to reality (sim2real).

Defining the boundaries of this focus area is challenging due to the lack of a fully-formed consensus on the definition of AGI

**19%** of promising research in this area is related to brain and mind simulation

The evolution of the area is defined by a fundamental shift from building narrow, specialized models towards developing more general, adaptive, and autonomous systems

Growing public and regulatory pressure for ethical and reliable AI is a key driver for sustainable development of this area

A major challenge for this is given the demand for automation and autonomy, is ensuring the robustness and reliability of the developed technologies and systems

FOCUS AREA 9

# Human-machine interaction

# FOCUS AREA 9

# Human-machine interaction

## 1. Overview of the focus area

Human-Machine Interaction (HMI) is an interdisciplinary field focused on the research, design, and implementation of interfaces for communication and effective collaboration between humans and artificial intelligence systems. Its boundaries go beyond traditional graphical interfaces and cover all forms of two-way exchange of multimodal information, including speech, gestures, gaze, tactile sensations, as well as direct reading and decoding of signals of nervous system activity and stimulating effects on nervous tissue. An important aspect is the creation of tools for the integration and collective interaction of humans and AI.

There are currently three pillars within this area:

### 1. Technical means of direct interaction with the human nervous system

This subfield focuses on systems that interact with the human nervous system through direct contact. It includes the development of brain signal reading systems, as well as the creation of specialized hardware and low-level software that support bidirectional interaction with neural tissue. Particular attention is paid to the creation of biocompatible means for reading neural population activity signals with maximum detail and coverage. Combined with modern artificial intelligence technologies, the technical solutions developed in this subfield will not only revolutionize medical rehabilitation but also form the basis for the full integration of natural and machine intelligence.

### 2. Technical means of traditional human-machine interaction

This includes the development of immersive environments (virtual, mixed, and real) with multimodal interaction, as well as technical means for delivering feedback through natural sensory channels (visual, auditory, tactile, and olfactory). This subfield also explores and develops tools and algorithms for processing and interpreting human actions through gesture, emotion, and speech recognition, as well as user intention prediction and adaptive interfaces. The formation of a comprehensive assessment system using VR/AR testing environments, multi-criteria indicators (accuracy, delays, fatigue), physiological markers and confidentiality verification methods will allow an objective assessment of the effectiveness and security of interaction systems.

### 3. Methods and algorithms of human interaction

This sub-area is dedicated to developing a methodology and algorithmic component to support the joint work of humans and AI within the technical frameworks of the first two areas. This involves the creation of technologies to expand human capabilities and facilitate effective collaboration in human-machine teams, collaborative robots, and decision-making support. A separate, important component is research into deciphering the «brain code» and interpreting brain activity signals in invasive and non-invasive bidirectional brain-computer interfaces. Particular emphasis is placed on machine learning and adaptive systems capable of self-tuning and personalizing interactions. A key aspect is the creation of a comprehensive system of multi-criteria metrics (accuracy, latency, fatigue), physiological markers, and privacy verification methods, which will enable objective assessment of the effectiveness and security of interaction systems and ensure the progressive development of human-machine interaction systems.

The evolution of human-machine interaction has moved from interactive text interfaces (command line interface, CLI) that require knowledge of commands to intuitive graphical user interfaces (GUI). The key transition was made by the development of Xerox PARC, popularized by the Apple Macintosh, which introduced the WYSIWYG paradigm and made the mouse manipulator the main means of navigation. The next revolution was the widespread introduction of multi-touch touchscreens, the principle of which was developed back in the 1970s, but became widespread thanks to the iPhone, defining the modern Touch User

Special thanks for work on this focus area to:
A. Ossadtchi, A. Savchenko, I. Lavrov, C. Guger

Interface (TUI) standard. The improvement of speech recognition technologies and the development of LLM, which serve as the algorithmic basis of modern assistants, not only opened up the possibility of natural voice communication through VUI, but also radically increased the effectiveness of traditional CLI interaction, especially when programming and working with operating systems.

Recent progress in the technology of direct interaction with the brain using brain-computer interfaces (hereinafter — BCI) It provided decoding of not only motor intentions, but also speech from brain activity, as well as the possibility of direct modulation of neural processes, which, on the one hand, can bring information exchange between humans and AI to a qualitatively new level, but on the other hand entails ethical risks associated with unauthorized reading of semantic content and direct effects on the user's brain.

The following 5 events can be identified that have had the greatest impact on the development of the field of human-machine interaction over the past 5 years:

- **The explosive growth and accessibility of generative AI,** in particular large language models (2022-2023), has radically changed the very principle of interaction with a machine — from text input and clicks to natural conversation. Voice assistants (Alexa, Siri) have received a powerful upgrade, which has made speech interaction much more meaningful and useful.
- **The creation of invasive speech BCIs (2024-2025)** operating at an almost natural speed has highlighted the potential of direct computer interaction technology, including the ability to decode semantic, not just motor speech contexts.
- **The launch of Neuralink into human clinical trials (2024),** namely the beginning of experiments with implanting a chip into the human brain, was the key to attracting great attention from the public and investors to invasive BCI.
- **The development of the metaverse and augmented reality (XR), the legitimization of spatial computing, and the success of noninvasive BSI.** The announcement of Meta and the «metaverse» (2021) launched a technology race for immersive interaction, which stimulated the development of VR/AR headsets (Meta Quest, Apple Vision Pro) with improved systems for tracking gaze, gestures and facial expressions, as well as haptic feedback technologies and the decoding of electromyographic activity (Ctrl-labs) necessary for interacting with digital objects embedded in the user's physical world.
- **Consolidating the principles of AI Ethics (2019-2023):** Initiatives by the OECD, the EU and others have consolidated transparency, fairness and accountability as the foundations of trust in AI, stimulating research at XAI. Launching initiatives on AI Regulation (EU AI Act, 2021–2024): the creation of a legal framework directly shapes HMI's priorities, requiring a focus on robustness, transparency, and human oversight. This is especially important in the context of the COVID-19 pandemic (2020–2022), which acted as a catalyst for the introduction of remote work tools and the virtualization of human presence.

As the capabilities of AI and computing systems grow, it is the HMI that becomes the bottleneck. Without radically improving interfaces — increasing their speed, reliability, and naturalness — humanity will not be able to fully exploit the potential of AI. In the long term, HMI opens the way to fundamentally new possibilities, from "computational extenders" of human cognitive abilities to medical neuroprostheses that require a natural exchange of information with the brain. However, the rapid development of this field also creates serious risks associated with neuroprivacy, data protection, and technology abuse. Therefore, the development of accountable and explicable HMI systems is not only a technological, but also a strategic priority on which national security and social stability depend.

## 2. Overview of current developments

The modern development of human-machine interaction (HMI) is characterized by the convergence of AI, robotics, and neurotechnology, accompanied by a fundamental revision of interaction paradigms. Generative AI and LLM have revolutionized natural language interaction between humans and machines. There is a consistent transition from interfaces that require special training to systems based on real — world models, and further to proactive multimodal interfaces that can anticipate user intentions.

In the field of neurotechnology, progress is particularly noticeable: modern speech IMCS have achieved decoding of neural signals at an almost natural rate,

and bionic prosthetics demonstrate success in creating feedback interfaces. Preclinical practice has been enriched with the first Neuralink implants, and the clinic already uses adaptive deep brain stimulation systems to restore motor function in patients with Parkinsonism. Special attention is being paid to the development of technical means for reading brain activity with high spatial and temporal resolution, devices capable of recording cortical activity with a density of 44 contacts per square millimeter have been demonstrated, the first tests of such systems on humans are being conducted (Precision neuroscience Inc.) and AI algorithms for decoding such signals are being developed. Nanoparticle-based technologies for creating brain contact open up the possibility of non-invasive and high-precision modulation of brain activity, which, in turn and in combination with other trends, raises a number of ethical challenges that must be overcome for the sustainable development of the HMI field.

Explicable AI (XAI), which forms the algorithmic basis for trusted human-machine interaction, is of particular importance in this process. The explainability of AI decisions is transformed from an additional function into the basic principle of designing HMI systems, providing transparency of decision-making processes and interactive opportunities for joint AI and human reasoning. including visual support for the logical inference process.

Breakthroughs in brain-computer interfaces (BCI):

- Speech BCIs Demonstrate the decoding of neural signals into speech at an almost natural rate.
- Bionic prosthetics: Successful implementation of prostheses with sensitivity, both using direct interaction with the cortex and controlled by peripheral electromyographic signals and carried out through stimulation of peripheral nerves.
- The first patients with a Neuralink implant and public demonstrations of cursor/input control. The fact of the first implantation and stable pointer/typing control has increased interest in fully implantable wireless BCIs (including raised issues of safety, signal stability and rehabilitation scenarios, etc.).
- A separate category of OTC hearing aids. The creation of a market for hearing assistants has dramatically lowered access barriers and stimulated innovation in user audio interfaces (self-tuning, integration with mobile devices), setting a precedent for consumer neuro- and sensory devices.
- Synchron's first endovascular BCI in the USA as a new "class" of minimally invasive interfaces, actually a bridge between non-invasive EEG and cortical arrays.

## 3. Research challenges shaping and limiting the focus area development

The rapid development of the HMI field creates a number of challenges that research efforts should address today.

### CHALLENGE 9.1

#### Achieving sustainable robust and synchronized multimodal fusion

It is critically important to make this merger robust and able to take into account the social context and user intentions in multi-user HMI scenarios. Effective integration of asynchronous and heterogeneous data (speech, vision, gestures) is required no conflicts. The urgency of the challenge is reinforced by the development of multimodal models (Vision Language Models, VLM). Failure to synchronize leads to the creation of unreliable interfaces, and the degree of unreliability increases exponentially with the number of potentially conflicting interface modalities. This limits the use of AI in critical and sensitive areas. Successfully overcoming this challenge allows you to create "invisible" interfaces and proactive, anticipatory interfaces that are able to understand the user and work ahead of time.

### CHALLENGE 9.2

#### Overcoming the biocompatibility and long-term stability of invasive implants

The body's immune response (gliosis) to the implant leads to degradation of the quality of the neural signal over time, requiring repeated replacement operations or making long-term use impossible. The main technological obstacle to the commercialization and widespread clinical use of chronic invasive BCI makes therapy potentially dangerous and economically impractical. Today, research is actively conducted on new materials (flexible electronics, hydrogels), bioinert coatings, and miniaturization of electrodes to reduce the immune response. Synaptic neural interfaces are developing, in which contact with the nervous tissue is established due to the formation of a natural synapse on the electrode site.

### CHALLENGE 9.3

#### Decoding semantic intent versus motor commands

Most successful IMCS decode motor commands (the intention to move a hand). It is much more difficult to decode abstract thoughts, inner speech (intention), or an emotional state directly, without relying on motor correlates. This limits the use of BMI to restore higher cognitive functions (for example, communication in completely paralyzed patients) and to create next-generation interfaces. Today, LLMs are used to solve this problem for the contextual interpretation of neural signals, research in the field of decoding classroom and visual imagination (for example, speech IMCS).

### CHALLENGE 9.4

#### The study of coding principles for shaping the effects on the brain

The challenge is to decipher the fundamental "language" of the brain — the neural code — which converts information into patterns of electrical and chemical activity. Solving this problem will allow us to move from simple observation of brain activity to targeted programming of neural ensembles, opening the way to the creation of fundamentally new systems: from medical neuroprostheses that restore lost functions (vision, hearing, movement) to direct brain-computer interfaces for controlling complex systems with the power of thought.

### CHALLENGE 9.5

#### Overcoming administrative and legal barriers for clinical trials

Strict regulatory requirements and ethics committees make it difficult to conduct research with invasive BCI in humans. An additional barrier is the lack of established practice and infrastructure for such risky experiments with patients who need them most. The current situation slows down the pace of research and technology transfer from the laboratory to the clinic. Patients are being denied access to potentially breakthrough therapies. The world is practicing the creation of "hub" clinics at leading universities (for example, Mass General Brigham at Harvard), the development of accelerated regulatory procedures for medical devices of the "breakthrough therapy" category, and the involvement of patient communities in research design.

## 4. Long-term research tasks

### ★ TASK 9.1

#### Creation of intuitive agents that understand the user's requests and expectations, his emotional state, etc.

Unlike classical interfaces, interaction with such systems is based on the principle of goal setting — when a person sets not an action, but the desired result, and the agent chooses the optimal way to achieve it. For this purpose, multimodal models of speech perception, gestures, gaze, and physiological reactions are used, as well as architectures such as Vision-Language-Action and memory-augmented agents that can take into account the context, social roles, and emotional state of the user. Such systems can significantly reduce cognitive load and interaction time, which is especially important when controlling robots, autonomous systems and complex digital environments. It is promising to design systems that support the networking of mixed teams (people + AI agents) with a dynamic distribution of initiative, develop models to take into account the social context and algorithms for effective collective thinking.

### ★ TASK 9.2

#### Research and development of bi-directional brain-computer interfaces, decoding the "brain code" and the creation of fundamental data models for functional brain mapping

Creating a closed-loop "intention-machine action-feedback" system using multimodal brain activity recording methods, adaptive decoding based on foundation models, and reinforcement learning algorithms for real-time adjustment. Key requirements include ensuring minimal latency (<100 ms), high decoding accuracy, and increased speed of information exchange in non-invasive systems.

### ★ TASK 9.3

#### Establishment of a metrological base for HMI assessment

Today, there are no uniform standards for measuring parameters such as user trust, cognitive load, engagement, or the level of agent autonomy. The formation of a set of metrics and test protocols will allow an objective assessment of the effectiveness and safety of interaction systems. VR/AR sandboxes, multi-criteria metrics (accuracy, latency, fatigue), physiological

indicators, and privacy testing methods are expected to be used as tools. The development of an open HMI benchmark and certification criteria will create the basis for comparability of solutions, accelerate the launch of innovative products to the market, and increase trust in human-AI interaction technologies.

## 5. Important takeaways // Expert opinion

Human-machine interaction is becoming a strategic bottleneck of digital transformation: the quality, speed of information transfer, the naturalness of such interfaces and the degree of trust in them determine to what extent the potential of AI will be realized. This applies both to systems that ensure conscious human-machine interaction, and to neural interfaces that enable direct contact with the brain and nervous tissue and serve to replace lost functions in patient users or augment them in healthy users.

The trajectory of development of traditional means of human-machine interaction is shifting towards multimodality and neuro-cognitive systems: the combination of text, speech, vision and gesture has actually become the standard. There is a development of «invisible» interfaces and smart environments (Ambient/IoT), providing contextual and unobtrusive interaction. An important trend is the transition from the "one user, one machine" model to multi-user scenarios, human-AI partnerships, and machine-to-machine communication with human moderation.

Pinpoint breakthroughs have been recorded in neural interfaces: invasive speech decoding at speeds close to natural; successful attempts at semantic decoding, prosthetic limbs with sensory feedback through peripheral stimulation with EMG control; the use of BMI-mediated spinal cord stimulation; commercial availability of condition-dependent RNS systems for the treatment of epilepsy, rigidity and tremor; demonstration the information content of ultra-high-density EEG (about 4K channels).

Nevertheless, the main source of information about the activity of nervous tissue, providing the necessary amount of information, is so far invasive electrode systems, ideally providing access to the activity of a large number (more than 1000) of individual neurons. For the sustainable development of this area, it is necessary to create such implantable electrode systems with long-term biocompatibility and recording the activity of tens of thousands of neurons in spatially distributed areas of the cerebral cortex. An interesting direction is the formation of contact with the nervous tissue due to the synaptic interface.

It is essential to use the interface with the nervous tissue not only for reading information, but also for stimulation, which is necessary to close the feedback loop (tactile / sensory) in order to provide a sense of agency, increase the functionality of prosthetics and reduce cognitive load when using them. To solve this problem, it is critically important to use complex context-dependent neural network coding algorithms, distributed in space and time patterns of stimulation, providing the most natural feedback. Optogenetic and thermogenetic technologies are also used for high-precision and spatially selective stimulation of nervous tissue. In addition, nanoparticles are currently being considered as an alternative minimally invasive means of forming bidirectional contact with nervous tissue.

The most important organizational aspect determining the competitiveness and the possibility of developing neurointerface technologies is the creation of a domestic and international legal framework legitimizing experiments to replace lost functions in real patients using prototypes of neurointerface systems.

The path of development of human-machine interaction technologies lies in the direction of creating holistic, trusting and adaptive ecosystems. The key features of these ecosystems will be their ability to synergize with humans, from dynamic interfaces that anticipate intentions to neurocognitive technologies that provide deep integration at the biological level. Further progress will be determined not only by technological breakthroughs, but also by the successful solution of complex tasks at the intersection of regulation, ethics and cybersecurity. Creating an environment where technology not only provides tools, but also enhances human potential in partnership is a central task on the way to realizing the full range of digital transformation opportunities.

The strategic bottleneck of digital transformation: it is the quality, speed, and naturalness of interfaces that determine the extent to which the potential of AI will be realized

The most important challenge of the direction is the development of models that understand the social context, roles and hierarchies

Breakthroughs in the field of Generative AI and LLM have significantly changed the paradigm of machine-human interaction and the research landscape of the entire field

An important challenge for the direction, taking into account the demand for automation and autonomy, is to ensure the robustness and reliability of the technologies and systems being developed

**80%** tasks in one way or another, they are related to the expansion of human capabilities through various kinds of collaboration with AI

FOCUS AREA 10

# Society in the AI era

# FOCUS AREA 10

# Society in the AI era

## 1. Overview of the focus area

Along with benefits provided by AI technologies, we are seeing their comprehensive transformational impact on social institutions. This impact is cross-cutting in nature as it affects various areas of human activity. In the economic sphere, this is reflected in structural transformation of conventional industry sectors and the emergence of fundamentally new sectors, the so-called "industries of the future". In the social sphere, this involves a profound reshaping of the way of life, and transformation of employment models and educational paradigms. The cultural sphere is witnessing a change of content generation and distribution modes, which leads to the rethinking of the very foundations of cultural production. In the sphere of science and technology, AI gives rise to new research paradigms, marking the beginning of a new era in the organization of scientific knowledge.

*"In the coming years, the influence of AI will drive a new evolution, which highlights the need for the implementation of AI algorithms to improve the standard of living and develop in parallel with AI",* commented Professor Azidine Guezzaz, associate professor of computer science and mathematics at Essaouira Higher School of Technology at Cadi Ayyad University.

As part of collective discussions, the expert community has repeatedly emphasized that technological breakthroughs entail major social and ethical risks and AI adoption often outpaces the development of a supportive regulatory framework for the technology both on the national level and internationally. Researchers have pointed out: *"We are creating technologies without waiting for the emergence of a 'global social contract' regarding their acceptability".*

This focus area conceptualizes interaction between artificial intelligence and society as a three-way relationship covering governance, ethics, and social and economic transformation, and emphasizes that these areas mutually strengthen each other and should be developed in unity in order to support human-centric development. Thus, following international foresight sessions, the following core subareas have been singled out within this focus area:

### 10.1 Global AI governance mechanisms, including AI regulation:

- **National and global AI governance systems.** In recent years, most jurisdictions have been concerned with building an AI governance system both on the national level and globally. The fact that AI governance should transcend national borders and form a global framework based on multilateral cooperation has been repeatedly pointed out and incorporated into practice. Despite significant variations in national approaches to AI regulation, further development of the AI governance system necessitates establishing a more multifaceted and comprehensive paradigm that goes beyond the narrow limits of national regulatory models. At the same time, many researchers point out a systemic lag between the development of the legal framework and the pace of technological progress, which creates a regulatory vacuum and, consequently, intensifies public concern and social wariness. Nevertheless, it is also noteworthy that international cooperation is expanding in the establishment of regulatory sandboxes for AI on the national level, from Brazil to Mozambique and Indonesia, as well as across the European continent, helping governments and entrepreneurial ecosystems find an optimal balance between AI regulation and innovation (in the near future, the European Union plans to issue the relevant guidelines on their implementation).

- **International cooperation.** The past year has seen a tendency towards international cooperation on AI. Researchers emphasize that it is critically important to establish a broad international dialogue and consensus on AI governance and development in order to promote safe, responsible and trustworthy AI technologies. Everyone is working towards this goal, with initiatives ranging from the efforts

Special thanks for work on this focus area to:
S. Belouali, A. Neznamov, P. Chen, J. Pita Costa

of international organizations, such as the UN General Assembly, which has provided a platform for the Global Dialogue on AI Governance and is working to establish a scientific panel, where countries will convene annually to formulate rules for the development of safe and accountable AI systems, to calls to action on the national level. In the latter case, special mention should be made of China's initiative to establish the World Artificial Intelligence Cooperation Organization, where all countries would be able to participate on equal terms and cooperate in the formulation of international rules for the development of AI technologies.

### 10.2 AI ethics

Ethical issues are the central focus of discussions about society in the AI era. Sustainable development of technologies is only possible in compliance with the principles of non-discrimination, transparency and explainability of algorithms, data protection, safety and reliability, accountability and control. The expert community has come to an understanding on the need for "ethics by design", whereby these guidelines are embedded in a system at the development stage. This is especially important amid growing focus on the risks of bias and unfairness, which may be perpetuated and exacerbated through AI technologies. AI ethics should not be confined to procedural concepts of fairness and transparency; rather, it should also include social and technical aspects of ethics, covering such matters as algorithmic fairness, epistemic diversity and inequitable power structures embedded in data collection and system design processes.

### 10.3 Studying the impacts of AI technologies on society

- **Economy and labor market.** In the economic sphere, AI technologies are a powerful driver that can improve productivity and transform the labor market. However, at the same time, there is a risk of growing social disparity, job replacement, and the concentration of capital in large corporations. As pointed out by the scientific community, *«AI accelerates production but does not necessarily lead to a fair distribution of benefits».* As part of international discussions, the question of fair profit distribution is raised and compensatory measures are proposed, such as automation taxes, redistribution of benefits, and even the concept of universal basic income as a potential tool for mitigating social impacts. Economic transformation driven by AI should be analyzed from the perspective of the political economy of technology, given how automation redistributes value, rights over data and bargaining power, and sparks off new debate on dividends produced by AI, the concept of data as labor, and fair mechanisms for redistributing created value.

- **Cultural identity.** In terms of the cultural aspect of development of AI technologies, special focus should be given to the principles of cultural identity underlying AI technologies: the language and style of communication; AI training on historical and regional samples; resisting cultural standardization; models of politeness and "taboos"; political and social contexts. Experts emphasize the need to preserve cultural diversity, promote digital literacy and a critical view of AI outputs. Otherwise, society is running the risk of homogenization of cultural codes and the rise of the phenomenon of "post-truth". This aspect can be reinterpreted in the context of digital sovereignty and pluralism, which highlights the need for linguistic and epistemic diversity in AI development in order to prevent algorithmic homogenization and preserve the cultural heritage in the digital era.

- **Examining the boundaries of acceptability of AI autonomy to people in various spheres.** Examining the boundaries of acceptability of AI autonomy to people as users of AI systems in various social contexts is an important aspect, as this parameter determines the willingness of society to delegate decision-making in such areas as healthcare, transportation, financial services, etc. to systems. This research should be focused on a comprehensive analysis of the interplay between technical capabilities of algorithms, psychological factors such as transparency and trust, as well as contextual variables such as the risk level and cultural characteristics.

Professor Dr. Suyanto, Rector of Telkom University (Indonesia), has pointed out: *"Research should be aimed at preserving the sovereignty of data, individuals and culture in the era of 'post-truth' ".*

The history of interaction between society and technology demonstrates that each instance of large-scale automation not only resulted in productivity growth but also posed major social challenges. Industrialization in the 19th century gave rise to the Luddite movement; robotic automation in the second half of the 20th century sparked off debate on employment, and the digital revolution in the early 21st century has raised the issues of privacy and control over data. These examples show that technological progress invariably requires society

to adapt by developing social, legal and cultural mechanisms that can mitigate conflicts.

The current stage in the development of AI technologies also follows this pattern. The first solutions were implemented locally and selectively, gradually transforming entire industries. The widespread adoption of machine learning systems and neural networks has not only unlocked the potential for speeding up processes and boosting performance, but has also created new risks, ranging from discrimination in algorithms to privacy and transparency threats, from job replacement to the restructuring of the labor market, from the homogenization of cultural codes to the development of a new identity. In response to these challenges, codes of ethics, rules of "ethics by design", standards for assessing the impacts of AI and regulatory frameworks have started to emerge; however, they have all been developed after the event, which confirms that there is a consistent tendency for regulation to lag behind technological progress.

As part of scientific discussions, experts have expressed an opinion that the future should be shaped not by replacing humans but by enhancing their capabilities.

Thus, the history of automation and adoption of AI technologies reflects a repeating pattern: a technological leap opens up new horizons but at the same time raises the issues of social fairness, regulation and cultural adaptation, and the need for international cooperation.

Key social and economic developments that have made the biggest impact on the shaping and development of this focus area include the following:

- **Widespread commercialization and integration of generative AI models** into publicly available services and consumer products that we have been witnessing in recent years is creating a fundamentally new technological paradigm. Despite the potential for optimizing user experience and expanding functional capabilities, this development has given rise to a set of major challenges that go beyond purely technical tasks. These include ethical dilemmas related to content authorship, the fundamental issue of AI "hallucinations" (when models generate plausible but factually inaccurate information) and the risk of biased decision-making.

- **Risks posed by AI technologies** necessitate developing approaches to the regulation of AI technologies both on the national level and internationally:
  - **Worldwide adoption of ethics documents.** The past few years have seen a shift from private initiatives undertaken by individual companies to the development of ethical frameworks at the level of industry associations, national strategies and international organizations. In the early 2020s, the focus was on corporate principles of "responsible AI", whereas by 2021–2023, a number of documents were adopted by governments and intergovernmental entities. These include the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021), the Code of Ethics in the Field of Artificial Intelligence in Russia (2021), the OECD and G20 principles, etc. Over the past five years, there has been a shift towards formalizing ethical principles in national strategies and developing global frameworks, which reflects a higher level of maturity of the debate and attempts to translate values into political and legal mechanisms.
  - **Acceleration of legislative initiatives.** Between 2020 and 2025, there was a surge in the number of legislative initiatives aimed at regulating the use of AI. According to the 2025 Stanford AI Index, legislative mentions of AI rose by 21.3% compared to 2023. The global community is gradually progressing to the development of statutory regulation models for AI technologies. Currently, there are restrictive, hybrid and pro-innovation approaches to AI regulation.
- **Digital divide between developed and developing countries.** Lack of equal access to AI technologies hinders competition both on the national level and internationally. The dominance of major players on the AI market has a negative impact on overall global development in the sphere of AI. Leading economies continue to introduce AI in industry, public administration and the educational system, boosting productivity and efficiency, whereas developing countries are at risk of falling behind even more. The resulting imbalance exacerbates global inequality in terms of income levels, the quality of education and innovation potential, creating a perception of digital colonization, where developing countries are excluded from the development of AI technologies and merely serve as suppliers of data.
- **Economy and labor market.** Intensified debate over the impact of AI on the labor market has become one of the highlights of the past few years.

The widespread adoption of generative models in 2022 and 2023 has accelerated the automation of intellectual tasks, sparking off an intense debate on the future of jobs. As part of scientific discussions, the participants of the foresight session repeatedly pointed out that hundreds of millions of jobs would be transformed or disappear, with new types of employment emerging simultaneously. The importance of this focus area is determined by the fact that it sets the boundaries of public trust in AI and defines the balance between innovation and risks. The ability of the government and social institutions to ensure transparency, accountability and a fair distribution of benefits is a prerequisite for social stability, institutional legitimacy and the sustainability of the labor market. If these matters are neglected, this may lead to a rise in discrimination, breach of confidentiality and growing social inequality, whereas systematic implementation of "ethics by design" and well-thought-out regulation will help minimize costs and provide a basis for long-term trust. In the sphere of economy, appropriate legal frameworks and socially responsible standards help reduce transaction costs, make implementation more predictable and reduce the risk of excessive concentration of market power. In the sphere of culture and education, the preservation of identity and diversity during AI adoption helps make the technology more socially acceptable, strengthens values and supports creative development. Overall, this focus area guarantees that the transition to the AI era will be guided not only by the logic of efficiency but also by the logic of sustainability, fairness and human dignity.

## 2. Overview of current developments

Many experts have expressed concern that the regulation of AI technologies lags behind AI adoption and fails to respond to negative consequences for society. The past few years have seen a rise in the number of regulatory initiatives focused on AI; however, all of them are based, in one way or another, on different fundamental approaches. For instance, the EU AI Act reflects a conservative approach to the regulation of AI technologies as it introduces a risk-based approach and imposes burdensome legislative requirements on developers of AI models deploying them in the EU. Another group of countries, including China and Russia, has opted for a more flexible approach whereby regulation combines incentives with selective statutory restrictions and self-regulation. Some countries, including the UK, Singapore, South Korea and Japan, are developing a pro-innovation approach to the regulation of AI technologies, which imposes minimal statutory restrictions and focuses on supporting scientific research and investment in AI.

During a scientific discussion, Liu Shu, Executive Secretary General of Shenzhen Association for Artificial Intelligence, expressed her expert opinion on AI regulation and compared approaches adopted by the European Union and China: *"The key to introducing effective regulation is its application. The European Union, which has placed special emphasis on safety, ethics and compliance of the technology, has imposed multiple restrictions, which has resulted in relatively slow progress in the use of AI. Meanwhile, China supports rapid development of the technology and protects key areas of its use by adopting the relevant legislation, which helps translate the technology into practical applications more quickly and supports widespread AI adoption".*

Despite different approaches adopted by countries at the national level, the AI governance system necessitates developing a more multifaceted and comprehensive approach to AI that goes beyond national regulatory frameworks.

At the same time, tools for ethical assessment of AI models are being actively developed and introduced both at the level of international organizations and national legislation of various countries and by individual leading corporations in order to detect and address bias and errors in AI algorithms, improve data protection mechanisms, develop human-centric AI models and build public trust.

The findings of the international foresight study have confirmed that various approaches to monitoring and evaluating AI systems are used in practice, leading to disagreement over matters related to control and accountability for the actions of AI systems.

Economic transformation triggered by the spread of artificial intelligence technologies is characterized not only by faster productivity growth but also by a noticeable redistribution of financial flows in favor of major corporations, which exacerbates social inequality. In response to these challenges, discussions have intensified on the global level concerning the search for mechanisms for the fair distribution of benefits, including introducing an automation tax and considering unconditional basic income models.

Simultaneously, in the cultural and social sphere, we are witnessing the rethinking of traditional paradigms influenced by AI: the notions of creativity, educational processes and even personal identity are being transformed. These changes highlight the issues of ethical responsibility of developers of algorithms and the urgent need to promote digital literacy among individuals, including the ability to critically assess AI outputs. To address these multifaceted problems, an interdisciplinary approach is necessary which would combine the competences of both technical specialists and experts in the field of humanities.

## 3. Research challenges stimulating or hindering the development of the focus area

The following research challenges stimulating or hindering the development of the focus area deserve a special mention:

### CHALLENGE 10.1

#### Developing an AI governance system and overcoming the digital divide in the sphere of AI

The development of a global artificial intelligence governance system is at an early stage, which is reflected in the wide variety of initiatives, legislative proposals and concepts put forward on the national level by international organizations and corporations. The key task is to consolidate these efforts in order to develop an equitable, sustainable and trustworthy model based on international cooperation. At the same time, experts emphasize that AI governance should aim to provide equal access to the technology and close the global digital divide. AI technologies should not exacerbate existing inequality; on the contrary, they should be a catalyst for development for all countries, including technologically vulnerable regions. Accordingly, the development of AI must be aligned with sustainable development goals and basic values of human society.

### CHALLENGE 10.2

#### International cooperation

The global community is gradually progressing to the development of internationally recognized standards for AI. The key task on this journey is to establish broad intergovernmental dialogue in order to achieve international consensus on AI regulation. It is especially important to progress from declarative statements to practical actions aimed at providing equal opportunities for all countries.

### CHALLENGE 10.3

#### Developing uniform standards for measuring ethical characteristics of AI models

Existing metrics, which form the basis of the methodology for ethical assessment of AI models, often fail to fully capture all aspects of compliance of AI models with ethical principles, from non-discrimination in AI algorithms and avoidance of accidental and undesirable correlations to accountability and control over the functioning of AI models. Accordingly, the expert community emphasizes the importance of developing comprehensive evaluation frameworks that will make it possible to compare AI models, thus providing a generally recognized evaluation standard that will establish a clear set of metrics required for compliance with ethical AI principles.

It is necessary to include the need for the development and institutionalization of methods for the assessment of ethical implications and human rights impact assessment (HRIA) in order to translate the "ethics by design" principle into a specific and verifiable practice. These tools support the effective incorporation of values at all stages in the life cycle of systems.

### CHALLENGE 10.4

#### Job loss and transformation of employment

The automation of a growing number of functions, including intellectual and creative tasks, causes serious concern over the future of the labor market. Many jobs are under threat of disappearing; at the same time, new forms of employment are emerging which require retraining. The research challenge consists in predicting these processes, and developing adaptation models and future-ready educational systems.

### CHALLENGE 10.5

#### Cultural pluralism and identity

Cultural and ethical characteristics of countries and regions vary significantly and may conflict with each other when developing universal international standards. This is reflected in language, communication styles, rules of politeness, historical and political contexts. The challenge consists in developing approaches that will make it possible to preserve cultural diversity when developing the minimum baseline universal standards for the global use of AI. Raising the importance of intercultural interaction to the level of human safety: large language models (LLMs) communicate specific

cultural representations that can threaten cultural identity. Research should establish how dataset filtering, annotation and design can help ensure that artificial intelligence will respect social relationships and collective dignity, not just individual dignity.

### ⚡ CHALLENGE 10.6

#### Impact of complex technologies, including AI agents, on human cognitive functions

New-generation AIs operating as autonomous agents pose heightened risks ranging from the loss of control over the behavior of systems to the increased opacity of their decisions. They reshape forms of communication, delegation of authority and the structure of social institutions. The challenge consists in examining long-term implications of the adoption of AI agents and developing frameworks that will make it possible to use them safely and for the benefit of society.

### ⚡ CHALLENGE 10.7

#### Examining the boundaries of acceptability of AI autonomy to people in various spheres

Rapid development of AI technologies and the gradual emergence of artificial general intelligence necessitates setting the limits for the delegation of decision-making functions to autonomous systems. This challenge is focused on comprehensive examination of social acceptability of various levels of AI autonomy across the entire range of vital areas of human activity. Research is focused on a set of social and psychological, ethical and cultural determinants of the attitudes of users of AI systems rather than on technological characteristics of the systems.

## 4. Long-term research tasks

### ★ TASK 10.1

#### Developing approaches to a global AI governance system

We are currently witnessing the gradual development of international cooperation on artificial intelligence. Approaches to global AI regulation are emerging which are aimed at turning the technology into a tool for addressing universal human problems rather than a source of new conflicts. In order to accomplish this goal, it is necessary to systematize data on national legal systems and on the cultural, social and political characteristics of various countries by developing standardized assessment tools. This will help provide an inclusive framework for a global agreement involving all countries, including the Global South. It is necessary to establish mechanisms for intercultural ethical audit in order to ensure the fairness of technologies being deployed in vulnerable regions, whose aim is not only to provide access but also to ensure the appropriateness of AI outputs in local contexts.

The deliverable for this task is a global agreement on AI and/or the establishment of an independent international body coordinating the functioning of the global AI governance system.

### ★ TASK 10.2

#### Developing national systems for AI regulation

In recent years, countries have been actively searching for effective models of AI regulation taking into account national characteristics. The development of regulatory frameworks is focused on cultural characteristics, strategic priorities and existing legal systems in each country. Defining "red lines", i.e. statutory limits on the use of AI reflecting societal values and ethical principles, is a major aspect of this process.

The deliverable is the development of well-balanced national regulation systems that both support the development of innovations and safeguard public interests. These systems are designed to seamlessly integrate national characteristics with international standards and provide a predictable and transparent environment for the development and adoption of AI technologies.

### ★ TASK 10.3

#### Developing uniform standards and metrics for social and ethical evaluation of AI and measuring social and economic consequences of introduction of AI technologies

To address this task, it is necessary to develop comparable metrics and transparent frameworks for evaluating non-discrimination, explainability, confidentiality, safety and accountability. Existing approaches are fragmented, which makes it difficult to compare systems and erodes trust. Measuring the impact on productivity, employment, income distribution and the concentration of market power. Developing scenarios of transformation of the labor market and mechanisms for the fair distribution of benefits.

The deliverable is an established universal evaluation system making it possible to compare various AI

models, improve transparency and make the adoption of the technology more manageable, provide a basis for certification, and develop validated guidelines for government policy, including regulatory and fiscal instruments, educational strategies and social support programs.

### ★ TASK 10.4

#### Examining the impact of AI technologies on society

New-generation AIs, including AI agents, are becoming increasingly autonomous, which increases the risk of opacity and uncontrollable behavior. These technologies can reshape models of communication, delegation of authority and the functioning of social institutions.

The key task is to define the component of human nature that we wish to preserve in the course of technology selection. This involves analyzing psychological and social implications of the delegation of decision-making and establishing the "right to refuse" in the course of interaction with AI systems.

The deliverables of this study, which consist in examining the boundaries of acceptability of AI autonomy to people, will make it possible to provide a scientific basis for developing human-centric interfaces and regulatory requirements maintaining a balance between the innovation potential of the technologies and the protection of fundamental rights and safety of individuals.

## 5. Important takeaways // Expert opinion

Given the transboundary nature of AI technologies, most experts and researchers agree that maintaining an international dialogue and developing a single global agreement on AI governance are the most challenging and top-priority task that needs to be addressed for the benefit of society and in order to support an even development of AI technologies.

It has been pointed out that the development of AI technologies is accompanied by the formalization of basic ethical requirements, such as transparency, accountability, data protection and non-discrimination. However, approaches to their implementation remain fragmented and vary from country to country, which makes it more difficult to compare practices and slows down the establishment of a single field of trust.

Social and economic implications are becoming increasingly obvious: automation leads to the transformation of the labor market and intensifies debate over the fair distribution of benefits. Taxation of automation, new forms of social support and the development of retraining systems are on the agenda.

The cultural dimension is becoming especially important. On the one hand, global technologies show a tendency towards standardization; on the other hand, there remains a need to respect cultural diversity, language and social norms. The level of social acceptability of AI is largely determined by public awareness and people's ability to critically assess the outputs of systems.

---

Mechanisms for global AI governance must be based on open global dialogue where AI regulation is viewed as a prerequisite for sustainable coexistence. The regulation system to be developed should have a socio-technical rather than purely technical nature, recognizing that artificial intelligence is a product of human choice integrated into the existing fields of power

---

Ethics should involve a process of collective comprehension rather than a set of rigid rules. This approach should incorporate the principle of "moral pluralism"

---

The problem of building a comprehensive AI regulation system lies not only in choosing the model (a restrictive or pro-innovation approach) but also in adopting a flexible regulatory framework capable of evolving in line with technical progress. This system should be underpinned by a minimum ethical consensus similar to the UNESCO consensus in order to gain public trust as a basis of legitimacy

---

# CONCLUSION

This study is intended as a "snapshot of the era", a large-scale effort attempting to consolidate knowledge about AI trajectories, limitations and windows of opportunity into a single corpus. We viewed AI not as a set of disparate technological trends but as an interconnected system encompassing architectures and algorithms, computations and energy, data and law, foundation and generative models, safety and trust, specialized solutions and multi-agent systems, elements of reasoning on the path to AGI, human–machine interfaces, as well as social and economic impacts and regulatory aspects — ten topics forming part of a single landscape. This approach helps to see systemic cause-and-effect relationships: what hinders or accelerates development and where the boundaries of scalability will be in the coming years.

In terms of methodology, this work was underpinned by broad international cooperation: 21 foresight sessions and interviews with more than a hundred leading researchers from 36 countries, supplemented by an analysis of open-source information and industry reviews. This combination of an academic perspective and implementation practice has made it possible to outline both consensus areas (e.g. The critical importance of energy consumption and data shortage) and fundamental disagreements (e.g. about the boundaries of agency, the path to explainability, architecture priorities) and, most importantly, to identify "bottlenecks" in the engineering cycle for AI. As a result, instead of a static "snapshot", we have produced a dynamic map that can be updated on a regular basis and is comparable over time.

Based on the findings of the large-scale study that we have conducted, we conclude our report with three key takeaways:

**Takeaway 1.**

The trajectory of AI quality is currently determined by the co-optimization of algorithms, software and hardware amid growing computing and energy costs rather than by a single "magic button".

**Takeaway 2.**

The era of agentic and hybrid systems is beginning, marked by a shift from passive assistants to proactive AI agents and to architectures that combine statistical learning with explicit knowledge and models of the world.

**Takeaway 3.**

Steady progress is hampered by a "double shortage", namely a shortage of energy and high-quality data; synthetic data, active and in-context learning and new sources of data from dynamic environments are a necessity rather than a luxury.

The practical usefulness of this study consists in the fact that it offers a common language for researchers, engineers, businesses and regulators. We have sought to avoid suggesting one-size-fits-all solutions; instead, we have outlined the boundaries of various solutions, showing where large-scale foundation approaches are suitable and where local adaptation and edge AI are critical; when data centralization is appropriate and when federated learning is preferable; how to balance gains in quality against the inference cost; how to measure trust and explainability for real-world human–AI interfaces. This framework will help design road maps in such a way that will ensure that they remain realistic and comparable across regions and industries.

We view this issue as an iteration of a "live" global AI monitor. In order to keep track of shifts, from changes in optimizers and compression techniques to surges in multi-agent systems and new classes of accelerators, monitoring needs to be done on a regular basis.

International cooperation in producing the report is a necessary prerequisite, as environmental and energy constraints, safety and privacy standards, the circulation of data and models, compatibility of tools and protocols are all transboundary phenomena by definition.

Collaboration between laboratories, companies and regulators helps reduce transaction costs associated with progress and speed up knowledge transfer from AI science into healthcare, education, industry and governance practices.

We would like to thank all participants of this research for a frank and professional discussion. This report provides a review of the work done to date, but it is by no means final. We invite our partners from various countries to join the "next round" in order to work together to maintain a shared field of knowledge, remove barriers, test hypotheses and ensure that the development of AI is responsible, economical and truly global.

# APPENDIX

## FOCUS AREA 1

### Architectures, machine learning algorithms, optimization and mathematics

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **1.1. Development of new machine learning algorithms** | 1.1.1 | Integration of machine learning with expert systems (hybrid AI) | | | | | | | | | | | |
| | 1.1.2 | Developing self-supervised learning methods | | | | | | | | | | | |
| | 1.1.3 | Developing universal methods for the processing of heterogeneous data structures | | | | | | | | | | | |
| | 1.1.4 | Developing algorithms for specialized hardware systems and computing devices | | | | | | | | | | | |
| | 1.1.5 | Developing in-context and verbal learning methods (verbal reinforcement learning; inductive learning; training of multi-agent systems; backpropagation) | | | | | | | | | | | |
| | 1.1.6 | Developing algorithms for multidimensional low-sample scenarios | | | | | | | | | | | |
| | 1.1.7 | Developing hybrid (conventional and quantum) algorithms, quantum neural networks | | | | | | | | | | | |
| | 1.1.8 | Developing new algorithms for domain-specific quality measurement of machine learning models | | | | | | | | | | | |
| | 1.1.9 | Developing self-evolving AI algorithms | | | | | | | | | | | |
| **1.2. AI architectures** | 1.2.1 | Developing adaptive methods for deep network architectures | | | | | | | | | | | |
| | 1.2.2 | Developing neural network architectures inspired by neurobiology and psychology, including spiking neural networks | | | | | | | | | | | |
| | 1.2.3 | Developing AutoML methods: meta-learning, automated feature engineering, hyperparameter optimization, automated model compression, etc. | | | | | | | | | | | |
| **1.3. Computation speedup** | 1.3.1 | Developing neural network compression methods: quantization, teacher-student, network pruning, etc. | | | | | | | | | | | |
| | 1.3.2 | Computation optimization for known neural network architectures (at the training and inference stages). Neural network structure and physics (including tensor and matrix networks) | | | | | | | | | | | |
| | 1.3.3 | Developing software tools for computation speedup | | | | | | | | | | | |
| **1.4. Distributed and federated learning** | 1.4.1 | Developing optimization methods for distributed and federated learning for large AI models: reducing "overhead costs" associated with data exchange, improving synchronization methods for distributed models, etc. | | | | | | | | | | | |
| | 1.4.2 | Developing distributed decentralized learning methods | | | | | | | | | | | |
| | 1.4.3 | Developing architectures (including mixed architectures) for federated learning | | | | | | | | | | | |
| | 1.4.4 | Preventing unauthorized access to data during processing and storage as part of federated learning | | | | | | | | | | | |
| **1.5. Mathematical foundation of AI** | 1.5.1 | Examining the mathematical foundations for reducing the complexity of machine learning models | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1.5.2 | Development of numerical methods of mathematical optimization for AI (including complexity theory) | | | | | | | | | | | |
| | 1.5.3 | Development of numerical linear algebra methods for AI | | | | | | | | | | | |
| | 1.5.4 | Study the foundations of stochastic methods for AI | | | | | | | | | | | |
| | 1.5.5 | Development and expanding information theory for AI | | | | | | | | | | | |
| | 1.5.6 | Research on approximation theory (explaining AI behavior) | | | | | | | | | | | |
| | 1.5.7 | Development of adversarial resilience (as applied to online learning) | | | | | | | | | | | |
| | 1.5.8 | Study the theoretical foundations for reinforcement learning and stochastic optimal control | | | | | | | | | | | |
| | 1.5.9 | Development of mathematically sound smaller deep learning models | | | | | | | | | | | |
| | 1.5.10 | Examination of the landscape of target functions and development more effective target functions and ways of learning | | | | | | | | | | | |
| | 1.5.11 | Study the theoretical foundations for generative models | | | | | | | | | | | |
| | 1.5.12 | Study the theoretical foundations for transformer models | | | | | | | | | | | |
| | 1.5.13 | Providing mathematical foundations for the understanding of multi-agent models | | | | | | | | | | | |
| | 1.5.14 | Study the theoretical foundations for the handling of uncertainty | | | | | | | | | | | |

## FOCUS AREA 2

## Computation for AI

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2.1. Development of specialized computing devices for AI (quantum, photonic, neuromorphic, etc.)** | 2.1.1 | Developing the architectures of specialized hardware and computing devices optimized for neural network architectures | | | | | | | | | | | |
| | 2.1.2 | Developing photonic processors and the relevant algorithms for the purposes of AI | | | | | | | | | | | |
| | 2.1.3 | Developing neuromorphic processors and the relevant algorithms for the purposes of AI; sensors, environment and actuators for neuromorphic processors | | | | | | | | | | | |
| | 2.1.4 | Studying and developing quantum processors and the relevant algorithms for the purposes of AI | | | | | | | | | | | |
| | 2.1.5 | Developing system software that will make the use of equipment more efficient | | | | | | | | | | | |
| | 2.1.6 | Developing methods and models for improving the efficiency of the learning process | | | | | | | | | | | |
| | 2.1.7 | Developing chip adaptation methods | | | | | | | | | | | |
| **2.2. Development of hardware and software suites for AI** | 2.2.1 | Building a high-speed network for data exchange between microprocessors | | | | | | | | | | | |
| | 2.2.2 | Developing new, more trusted libraries (using AI code generation and AI review of existing code according to the "strength of trust" criterion) | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2.3. Machine learning and AI frameworks** | 2.3.1 | Computation optimization, including on heterogeneous hardware | | | | | | | | | | | |
| | 2.3.2 | Frameworks for training and inference | | | | | | | | | | | |
| | 2.3.3 | Developing frameworks for symbolic and hybrid AI (e.g. SymbolicAI Framework, IBM Neuro-Symbolic AI Toolkit (NSTK), SmythOS, conventional symbolic AI tools (Prolog, expert systems, semantic networks/ontologies), neuro-symbolic hybrids) | | | | | | | | | | | |
| | 2.3.4 | Developing frameworks for agentic setups and apps (including embodied agents) (e.g. AutoGen, CrewAI, LangChain, LlamaIndex, Semantic Kernel, etc.) | | | | | | | | | | | |
| | 2.3.5 | Developing frameworks for prompt engineering (e.g. RTF, COSTAR, IEEI, APE, PECRA, OSCAR and TAG) | | | | | | | | | | | |

## FOCUS AREA 3

## Data for AI

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **3.1. Development of benchmarks for AI** | 3.1.1 | Development and maintaining benchmarks for comparing and evaluating the performance of neural networks of all formats (CNNs, transformers, LLMs, etc.) | | | | | | | | | | | |
| | 3.1.2 | Developing metrics for AI evaluation | | | | | | | | | | | |
| | 3.1.3 | Online evaluation of AI methods (real-world evaluation) | | | | | | | | | | | |
| | 3.1.4 | Assessing the effectiveness of safeguards | | | | | | | | | | | |
| **3.2. Data generation, conversion and maintenance** | 3.2.1 | Developing methods for building synthetic datasets | | | | | | | | | | | |
| | 3.2.2 | Active learning and crowdsourcing with assistance from expert communities, collecting subjective data from a large number of people | | | | | | | | | | | |
| | 3.2.3 | Studying data sequencing techniques (including curriculum learning) | | | | | | | | | | | |
| | 3.2.4 | Simulation systems for online data generation, model training and validation | | | | | | | | | | | |
| | 3.2.5 | Developing techniques for data quality assessment, filtering, curation and sorting | | | | | | | | | | | |
| | 3.2.6 | Developing data augmentation techniques | | | | | | | | | | | |
| | 3.2.7 | Building open datasets for the training of large-scale practical models | | | | | | | | | | | |
| | 3.2.8 | Developing verification methods and correction algorithms for reducing data drift and bias | | | | | | | | | | | |
| **3.3. Ensuring data privacy and security** | 3.3.1 | Developing differential privacy techniques: adding noise and distortion to data | | | | | | | | | | | |
| | 3.3.2 | Developing data anonymization techniques: making it impossible to recover confidential data | | | | | | | | | | | |
| | 3.3.3 | Developing methods for ensuring data confidentiality during federated learning | | | | | | | | | | | |
| | 3.3.4 | Watermarking personal and copyrighted data, and methods for detecting the use of such data in the training process | | | | | | | | | | | |
| | 3.3.5 | Watermarking AI-generated content, and methods for detecting such content | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

FOCUS AREA 4

## Foundation generative models

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4.1. Generative foundation models for character data** | 4.1.1 | Stydying and developing methods for the training and fine-tuning of generative foundation models | | | | | | | | | | | |
| | 4.1.2 | Studiing and developing methods for creating generative models (including RL in various applications) | | | | | | | | | | | |
| | 4.1.3 | Creating computationally efficient architectures for generative foundation models | | | | | | | | | | | |
| | 4.1.4 | Developing a high-quality data model based on generative learning models | | | | | | | | | | | |
| | 4.1.5 | Studying the impact of data quality on learning processes in generative foundation models | | | | | | | | | | | |
| | 4.1.6 | Developing methods for reducing the impact of hallucinations and quantifying uncertainty in generative foundation models | | | | | | | | | | | |
| | 4.1.7 | Studying modern architectures (including transformer architectures) for various sequential data processing tasks | | | | | | | | | | | |
| | 4.1.8 | Developing plausible generation techniques | | | | | | | | | | | |
| | 4.1.9 | Techniques for the implementation of reasoning in generative foundation models (taking into account various domains) | | | | | | | | | | | |
| | 4.1.10 | Developing representation learning methods | | | | | | | | | | | |
| **4.2. Generative foundation models for non-character data** | 4.2.1 | Developing generative foundation models for image and video processing | | | | | | | | | | | |
| | 4.2.2 | Developing generative foundation models for time series processing (sensors in robotics, lidar sensors, IoT sensors) | | | | | | | | | | | |
| | 4.2.3 | Studing, developing and using learnable representations of non-character data | | | | | | | | | | | |
| | 4.2.4 | Developing R&D focused on generative foundation models for 3D/spatial data | | | | | | | | | | | |
| | 4.2.5 | Developing generative foundation models for addressing tasks in the field of biology, pharmacology, meteorology and other areas of science | | | | | | | | | | | |
| | 4.2.6 | Developing effective architectures and methods for video understanding and processing, including VLM applications | | | | | | | | | | | |
| **4.3. Multimodal generative foundation models** | 4.3.1 | Developing mechanisms for augmenting foundation language models with capabilities for processing non-character modalities | | | | | | | | | | | |
| | 4.3.2 | Developing methods for the effective encoding of data in non-character modalities | | | | | | | | | | | |
| | 4.3.3 | Studying new methods for mixing various modality encoders | | | | | | | | | | | |
| | 4.3.4 | Development and study of multimodal generative models adapted for narrow domain-specific tasks | | | | | | | | | | | |
| **4.4. Knowledge transfer with adaptation of a generative foundation model** | 4.4.1 | Developing fine-tuning methods for generative foundation models (e.g. LoRA, P-tuning) | | | | | | | | | | | |
| | 4.4.2 | Developing methods for model distillation for narrow tasks | | | | | | | | | | | |
| | 4.4.3 | Developing techniques for personalized generation in various modalities | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **4.5. Augmentation of generative foundation models** | 4.5.1 | Studying and developing short- and long-term memory, as well as RAG and graph RAG | | | | | | | | | | | |
| | 4.5.2 | Studying and developing of in-context learning and prompt engineering techniques | | | | | | | | | | | |
| | 4.5.3 | Use of tools (e.g. API calls) in generative foundation models | | | | | | | | | | | |

## FOCUS AREA 5

## Safety, trust and explainability

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **5.1. Alignment** | 5.1.1 | Developing methods for mitigating risks associated with incorrect or malicious data (inaccurate data, restricted data, including information whose dissemination is prohibited by law, data that are not consistent with societal values) | | | | | | | | | | | |
| | 5.1.2 | Developing techniques for detecting algorithm footprints | | | | | | | | | | | |
| | 5.1.3 | Developing methods for incorporating values into general-purpose generative (language, multimodal) models (values of a society, a culture, individual groups), including the alignment of multiple preferences | | | | | | | | | | | |
| | 5.1.4 | Developing approaches to the standardization of development of benchmarks for measuring compliance with alignment goals; developing alignment benchmarks | | | | | | | | | | | |
| | 5.1.5 | Alignment of reasoning models | | | | | | | | | | | |
| | 5.1.6 | Multimodal alignment | | | | | | | | | | | |
| | 5.1.7 | Machine unlearning | | | | | | | | | | | |
| | 5.1.8 | Developing error correction techniques for AI models | | | | | | | | | | | |
| **5.2. Explainability of AI** | 5.2.1 | Formulating common approaches to ensure explainability and build trust in the operation of artificial intelligence: explainable AI (XAI) | | | | | | | | | | | |
| | 5.2.2 | Sudying and developing methods for post factum explanation (interpretation) of opaque AI methods based on neural networks | | | | | | | | | | | |
| | 5.2.3 | Developing interpretable and self-explainable models (self-explainable AI) | | | | | | | | | | | |
| | 5.2.4 | Developing argumentation and concept learning systems (concept & argumentation-based AI) | | | | | | | | | | | |
| | 5.2.5 | Studing and developing transparent AI methods based on formal systems, logic and knowledge bases | | | | | | | | | | | |
| | 5.2.6 | Developing explainability testing requirements and protocols for AI systems | | | | | | | | | | | |
| | 5.2.7 | Developing human-centered XAI, including aspects taking into account industry-specific characteristics | | | | | | | | | | | |
| | 5.2.8 | Ensuring the explainability of large language and multimodal models (LLM-/VLM-XAI) | | | | | | | | | | | |
| **5.3. Ensuring safe development and operation of AI** | 5.3.1 | Developing methods and infrastructure to ensure secure development of AI systems (MLSecOps) | | | | | | | | | | | |
| | 5.3.2 | Developing methods and software tools for protecting against attacks on datasets (including data poisoning and backdoors) | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 5.3.3 | Developing methods for finding vulnerabilities in code from third-party libraries used in the development of AI systems | | | | | | | | | | | |
| | 5.3.4 | Developing methods and software tools for protecting against attacks on AI models during training and execution | | | | | | | | | | | |
| | 5.3.5 | Developing security and trust benchmarks | | | | | | | | | | | |
| | 5.3.6 | Studying fundamental limits for security safeguards, including mathematical aspects taking into account empirical/heuristic methods | | | | | | | | | | | |
| | 5.3.7 | Developing methods for secure code base development using AI tools | | | | | | | | | | | |
| | 5.3.8 | Applying of federated learning methods to mitigate risks associated with the handling of sensitive data | | | | | | | | | | | |
| | 5.3.9 | Ensuring the security of LLM applications | | | | | | | | | | | |
| **5.4. Ensuring protection from the results of AI use for hacking purposes** | 5.4.1 | Developing methods (involving the use of AI) for detecting software vulnerabilities, ensuring cybersecurity, addressing social engineering, fabricated information, etc. | | | | | | | | | | | |
| | 5.4.2 | Developing methods for detecting and protecting against deepfakes, including watermarks | | | | | | | | | | | |
| | 5.4.3 | Developing methods for detecting attacks on cyberphysical systems, including autonomous driving | | | | | | | | | | | |
| | 5.4.4 | Developing benchmarks for deepfake detection techniques | | | | | | | | | | | |
| | 5.4.5 | Identifying and formalizing of best practices in addressing vulnerabilities in AI agents | | | | | | | | | | | |

## FOCUS AREA 6

## Narrow AI

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **6.1. Computer vision (CV)** | 6.1.1 | Studying and developing of effective training and execution methods for convolutional, transformer and hybrid architectures for CV tasks (including AutoML) | | | | | | | | | | | |
| | 6.1.2 | Developing foundation models for various CV tasks (such as SAM, DINOv2, CLIP, generative VLMs) (including addressing open-vocabulary classification, detection and segmentation tasks) | | | | | | | | | | | |
| | 6.1.3 | Developing computer vision techniques for the simulation of real-world scenarios (including embodied AI) | | | | | | | | | | | |
| | 6.1.4 | Developing fine-tuning techniques for specific computer vision tasks | | | | | | | | | | | |
| | 6.1.5 | Developing effective spatial representations (multimodal, multisensory, NeRF, Gaussian splatting, etc.) for computer vision tasks | | | | | | | | | | | |
| **6.2. Natural language processing (NLP)** | 6.2.1 | Studying and developing effective training and execution methods for natural language processing architectures (including AutoML) | | | | | | | | | | | |
| | 6.2.2 | Developing NLP for various programming languages | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **6.3. Other narrow AI technologies (S2T, RecSys, TSA, etc.)** | 6.3.1 | Studying and developing of effective training and execution methods for RecSys, S2T and TSA architectures (including automated learning) | | | | | | | | | | | |

## FOCUS AREA 7

## Control, decision-making, and agentic/ multi-agent systems

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **7.1. Development of reinforcement learning algorithms** | 7.1.1 | Conventional reinforcement learning (RL) development: distributional RL, risk-constrained RL, entropy-regularized RL, safe RL, etc. | | | | | | | | | | | |
| | 7.1.2 | Developing conventional reinforcement learning (RL): distributional RL, risk-constrained RL, entropy-regularized RL, safe RL, etc. | | | | | | | | | | | |
| | 7.1.3 | Developing RL policies in complex, dynamic environments: meta-learning, hierarchical RL, in-context RL, etc. | | | | | | | | | | | |
| | 7.1.4 | Developing offline-to-online RL and real-time learning | | | | | | | | | | | |
| | 7.1.5 | Developing RL involving the use of generative models and agents (agentic RL) | | | | | | | | | | | |
| | 7.1.6 | Developing methods for the transfer of learning between real-world and virtual environments | | | | | | | | | | | |
| | 7.1.7 | Developing self-evolving algorithms | | | | | | | | | | | |
| | 7.1.8 | Developing inverse RL algorithms | | | | | | | | | | | |
| | 7.1.9 | Developing algorithms for high-dimensional few-shot scenarios | | | | | | | | | | | |
| | 7.1.10 | Unsupervised reinforcement learning techniques | | | | | | | | | | | |
| **7.2. Agentic systems** | 7.2.1 | Developing universal action models of a physical agent: physical manipulations, multitasking | | | | | | | | | | | |
| | 7.2.2 | Developing universal multimodal models integrating text and other modalities: Vision-Language-Action (VLA) | | | | | | | | | | | |
| | 7.2.3 | Developing effective methods for agent knowledge acquisition through environmental interaction in order to achieve goals | | | | | | | | | | | |
| | 7.2.4 | Developing foundation agents for the virtual and physical world: multitasking, self-learning, operating in an open environment, with an open-ended list of tools | | | | | | | | | | | |
| | 7.2.5 | Developing learning methods involving the use of unlabeled behavioral data (learning without actions) | | | | | | | | | | | |
| | 7.2.6 | Developing automated agent design methods | | | | | | | | | | | |
| **7.3. Multi-agent systems** | 7.3.1 | Developing multi-agent architectures based on foundation models | | | | | | | | | | | |
| | 7.3.2 | Developing knowledge sharing capabilities for agents: learning from other agents and from people | | | | | | | | | | | |
| | 7.3.3 | Studying and developing a system of agents: a "master system" for building narrow agents, resource management, resilience, evolutionary selection, etc. | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 7.3.4 | Studying and developing virtual and physical multi-agent systems using various scenarios of interaction between agents for collective decision-making: competition, coordination, cooperation, hive | | | | | | | | | | | |
| | 7.3.5 | Studying the phenomenon of phase transitions in multi-agent systems as the number of agents increases | | | | | | | | | | | |
| | 7.3.6 | Automated design and topology optimization of multi-agent systems | | | | | | | | | | | |
| | 7.3.7 | Developing efficient environments for multi-agent systems (text, multimodal) | | | | | | | | | | | |
| | 7.3.8 | Creating hierarchical multi-agent systems (networks of networks and systems of systems) | | | | | | | | | | | |
| | 7.3.9 | Creating methods for scientific and engineering applications of multi-agent systems | | | | | | | | | | | |

## FOCUS AREA 8

## Elements of AGI

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **8.1. Reasoning and reflection** | 8.1.1 | Developing general reasoning models and decision diagrams | | | | | | | | | | | |
| | 8.1.2 | Development and use of the world-model and the self-model | | | | | | | | | | | |
| | 8.1.3 | Developing methods for and approaches to the modeling of self-awareness, self-reflection and self-criticism | | | | | | | | | | | |
| | 8.1.4 | Developing approaches to goal-setting and planning | | | | | | | | | | | |
| | 8.1.5 | Recognizing, analyzing and managing misleading behavior | | | | | | | | | | | |
| **8.2. Lifelong learning** | 8.2.1 | Developing active and passive learning methods | | | | | | | | | | | |
| | 8.2.2 | Developing online and offline learning methods | | | | | | | | | | | |
| | 8.2.3 | Developing learning methods that involve direct interaction with the environment or the use of preprocessed data | | | | | | | | | | | |
| | 8.2.4 | Analyzing the operation of models amid data instability (concept/feature drift) | | | | | | | | | | | |
| **8.3. Hybrid AI** | 8.3.1 | Developing universal mechanisms for extracting formalized knowledge and generating new knowledge | | | | | | | | | | | |
| | 8.3.2 | Conjunctive use of machine learning, symbolic AI and computer simulation tools | | | | | | | | | | | |
| | 8.3.3 | Intuition and emotions of AI models | | | | | | | | | | | |
| | 8.3.4 | Developing VLA (Vision-Language-Action) models | | | | | | | | | | | |
| | 8.3.5 | Evaluating AGI models | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **8.4. Embodiment** | 8.4.1 | Developing foundation models for managing a robotic system (cross-embodiment and cross-reality generalization) | | | | | | | | | | | |
| | 8.4.2 | Developing AI systems for autonomous robots | | | | | | | | | | | |
| | 8.4.3 | Developing multimodal and multisensory input technologies | | | | | | | | | | | |
| **8.5. Brain and mind simulation** | 8.5.1 | Modeling the functioning of the brain and the nervous system | | | | | | | | | | | |
| | 8.5.2 | Modeling human psychic processes | | | | | | | | | | | |
| | 8.5.3 | Creating twins of consciousness | | | | | | | | | | | |
| | 8.5.4 | Studying and developing memorization and forgetting mechanisms | | | | | | | | | | | |

## FOCUS AREA 9

## Human-machine interaction

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **9.1. Technical means of direct interaction with the human nervous system** | 9.1.1 | Research and development of bi-directional brain-computer interfaces | | | | | | | | | | | |
| | 9.1.2 | Developing new equipment and materials for invasive BCIs (high-density electrodes, biocompatibility) | | | | | | | | | | | |
| | 9.1.3 | Developing new equipment for non-invasive BCIs (wearable devices, ultra-high-density EEG, autonomic nervous system activity recording devices) | | | | | | | | | | | |
| | 9.1.4 | Developing technologies and methods for compression and transmission of signals of neuronal activity | | | | | | | | | | | |
| | 9.1.5 | Developing new ways of forming natural contact of nervous tissue with a cybernetic device (synaptic neuroyntheses, nanoparticles) | | | | | | | | | | | |
| | 9.1.6 | Co-design of hardware and software modules for multi-channel recording and stimulation of brain activity | | | | | | | | | | | |
| | 9.1.7 | Creating specialized microchips and software for processing bio-neural signals | | | | | | | | | | | |
| | 9.1.8 | Creating lightweight architectures and algorithms for real-time multimodal merging on devices | | | | | | | | | | | |
| **9.2. Technical means of traditional human-machine interaction** | 9.2.1 | Creating technologies for generative, adaptive and personalized human impact | | | | | | | | | | | |
| | 9.2.2 | Creating tools for forming teams of people and AI agents | | | | | | | | | | | |
| | 9.2.3 | Creating multimodal immersive environments to increase collaboration efficiency | | | | | | | | | | | |
| | 9.2.4 | Creating mobile devices for video presentation, vibro-tactile and olfactory feedback | | | | | | | | | | | |
| | 9.2.5 | Creating fundamental models to account for the social context in HMI | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **9.3. Methods and algorithms of human interaction** | 9.3.1 | Research of the "brain code" and creation of fundamental data models for functional mapping of the brain | | | | | | | | | | | |
| | 9.3.2 | Developing algorithms for effective functioning of human teams and AI agents (Human-Machine Teaming) | | | | | | | | | | | |
| | 9.3.3 | Developing algorithms for decoding brain activity in speech, motor, and visual neuroprostheses | | | | | | | | | | | |
| | 9.3.4 | Developing algorithms for decoding neuromyographic activity | | | | | | | | | | | |
| | 9.3.5 | Cognitive neuroprosthetics (memory restoration) | | | | | | | | | | | |
| | 9.3.6 | Research of effective and adaptive human-machine interfaces (HMI) in various modalities | | | | | | | | | | | |
| | 9.3.7 | Creation of intuitive agents that understand the user's requests and expectations, his emotional state, etc. | | | | | | | | | | | |
| | 9.3.8 | Multimodal machine interfaces. Immersive interaction in mixed reality | | | | | | | | | | | |
| | 9.3.9 | Expanding human capabilities through interaction with AI using brain-computer interfaces | | | | | | | | | | | |
| | 9.3.10 | Establishing a metrological base for HMI assessment | | | | | | | | | | | |

## FOCUS AREA 10

## Society in the AI era

| Subarea | | Research task | 2025 | 2026 | 2027 | 2028 | 2029 | 2030 | 2031 | 2032 | 2033 | 2034 | 2035+ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **10.1. Global AI governance mechanisms, including AI regulation** | 10.1.1 | Developing approaches to a global AI governance system | | | | | | | | | | | |
| | 10.1.2 | Developing national systems for AI regulation | | | | | | | | | | | |
| **10.2. AI ethics** | 10.2.1 | Developing and institutionalizing methods for the assessment of ethical implications and human rights impact assessment | | | | | | | | | | | |
| **10.3. Study of AI technology impacts on society** | 10.3.1 | Examining the impact of AI technologies on society | | | | | | | | | | | |

Year of expected breakthrough, i.e. the year when a major scientific achievement is expected to occur that can dramatically advance the accomplishment of the task in hand and make a significant impact on the development of the technology and the world as a whole.

Year when the task is exhausted, i.e. the year when the task can be viewed as accomplished and when no new fundamental or applied discoveries in this area are expected.

# AUTHORS OF THE FINAL REPORT

## The main editorial board

**Prof. Dr. Ajith Abraham**
Vice Chancellor, Sai University
**India**

**Dr. Arutyun Avetisyan**
Academician of the Russian Academy of Sciences, Director of the V.P. Ivannikov Institute for System Programming of the Russian Academy of Sciences
**Russia**

**Prof. Dr. Nebojša Bačanin-Džakula**
Full Professor/ Vice-Rector for Scientific Research/ Head of Applied AI Study Program, Singidunum University
**Serbia**

**Dr. Ricardo Baeza-Yates**
Full Professor at Universitat Pompeu Fabra
**Spain**

**Dr. Alexander Boukhanovsky**
Director of the MegaFaculty of Translational Information Technologies, Scientific Supervisor of the «Strong AI in Industry» Research Center
**Russia**

**Dr. Evgeny Burnaev**
Doctor of Physical and Mathematical Sciences, Director of the Artificial Intelligence Center at the Skolkovo Institute of Science and Technology (Skoltech)
**Russia**

**Prof. Ashraf Darwish**
Dean, Faculty of Computers and Artificial Intelligence, Obour University for Science and Technology (OUST)
**Egypt**

**Dr. Alexander Gasnikov**
Rector, Innopolis University
**Russia**

**Christoph Guger**
Chief Executive Officer of g.tec medical engineering GmbH
**Austria**

**Dr. Hou Haowen**
Assistant Professor at the Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), affiliated with Shenzhen University
**China**

**Andrey Kuznetsov**
Director of the FusionBrain Laboratory, AIRI Institute; Executive Director of Data Research, Sber
**Russia**

**Prof. Zheng Liang**
Vice Dean, The Institute for AI International Governance of Tsinghua University
**China**

**Dr. Andrei Neznamov**
Managing Director of the Human-Centered AI Center, Sberbank PJSC, Secretary-General, AI Alliance Network
**Russia**

**Dr. Ivan Oseledets**
Doctor of Physical and Mathematical Sciences, RAS Professor, AIRI Institute CEO
**Russia**

**Dr. Alexey Ossadtchi**
Director: Centre for Bioelectric Interface/Institute for Cognitive Neuroscience, HSE University
**Russia**

**Dr. Anh-Huy Phan**
Full Professor; Head, Laboratory of Intelligent Signal and Image Processing, Artificial Intelligence Center, Skoltech
**Vietnam**

**Dr. Anderson Rocha**
Full Professor, and Head of the Artificial Intelligence Lab., Recod.ai, Institute of Computing, Univ. of Campinas (Unicamp)
**Brazil**

**Dr. Ye Tian**
Professor, Anhui University
**China**

**Dr. Yuri Vizilter**
Director of the «AI and Computer Vision» Division at the State Research Institute of Aviation Systems, Scientific Director of the AI Institute at Moscow Institute of Physics and Technology, Doctor of Physical and Mathematical Sciences, Professor of the Russian Academy of Sciences
**Russia**

**Dmitry Yudin**
PhD, Principal Research Scientist at the Cognitive AI Systems Lab, AIRI, Head of the Intelligent Transport Lab at the Center for Cognitive Modeling, Institute of Artificial Intelligence, MIPT
**Russia**

### Dr. Rana Fayyaz Ahmad

Director of AITeC (Artificial Intelligence Technology Centre), National Centre for Physics (NCP), Islamabad

**Pakistan**

### Akmal Akhatov

DSc, professor, Vice Rector for international cooperation, Samarkand State University named after Sharof Rashidov

**Uzbekistan**

### Prof (Dr) Vijay Anant Athavale

Principal & Professor, Walchand Institute of Technology

**India**

### Prof. Chaim Baskin

Assistant Professor (Senior Lecturer) in the School of Electrical and Computer Engineering at Ben-Gurion University of the Negev; Head of the INSIGHT Lab and a member of the Data Science Research Center

**Israel**

### Prof. Saida Belouali

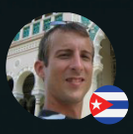
Professor, University Mohammed Premier (UMP)

**Morocco**

### Rustam Borovik

Section head of AI Transformation of the Human-Centered AI Center, Sberbank PJSC

**Russia**



### Dr. Milton García Borroto

PhD, Senior Researcher at the Center for Complex Systems of the Faculty of Physics of the University of Havana

**Cuba**

### Semyon Budyonnyy

Managing Director of the Advanced AI Technologies Development Department, Sberbank PJSC

**Russia**

### Nilolay Bushkov

Engineering Productivity R&D Architector, R&D Center, T-Technology

**Russia**

### Dra. Nayma Cepero

PhD, Full Professor. Faculty of Computer Engineering. Head of the Artificial Intelligence Research Group at Technological University of Havana. (CUJAE)

**Cuba**

### Viktoriia Chekalina

PhD, Senior Research Scientist, Multimodality research group, FusionBrain Lab, AIRI Institute

**Russia**

### Changsheng Chen

Professor, Shenzhen MSU-BIT University

**China**

### Dr. Peng Chen

Professor, Chair of Department of Intelligent Science and Technology, Anhui University

**China**

### Joao Pita Costa

Head of AI research at the International Research Centre on AI under the auspices of UNESCO - IRCAI

**UNESCO**

### Dr. Yudivian Almeida Cruz

PhD, data journalist, artificial intelligence researcher and university professor at the Faculty of Mathematics and Computer Science at the University of Havana

**Cuba**

### Denis Dimitrov

Managing Director of Data Research for the Kandinsky Base Models Department, Sberbank PJSC

**Russia**

### Prof. Olawande Daramola

Professor, Department of Informatics, University of Pretoria

**South Africa**

### Mukhamedieva Dilnoz

Doctor of Technical Sciences, Professor, Tashkent Institute of Irrigation and Agricultural Mechanization Engineers

**Uzbekistan**

### Egor Ershov

PhD, Head of Computational color photography research group, AIRI Institute

**Russia**

### Dra. Aylin Febles Estrada

PhD, Full Professor. Vice Minister of the Ministry of Communications

**Cuba**

### MSc. Rafael Luis Torralbas Ezpeleta

President of the Havana Science and Technology Park. Representative of Cuba to the Alliance for Artificial Intelligence (AI Alliance Network)

**Cuba**

### MSc. Héctor Rodríguez Figueredo

Vice-President of the Havana Science and Technology Park. Representative of Cuba to the Alliance for Artificial Intelligence (AI Alliance Network)

**Cuba**

### MSc. Allan Pierra Fuentes

Assistant Professor at the University of Informatic Sciences. Leader of Artificial Intelligence Projects at the Havana Science and Technology Park

**Cuba**

### Elizaveta Goncharova

PhD, Head of Multimodality research group, FusionBrain Lab, AIRI Institute

**Russia**

### Dra. Annet Morales González

PhD, Senior Researcher at the Center for Advanced Technology Applications, Pattern Recognition and Data Mining (CENATAV)

**Cuba**

### Azidine Guezzaz

Associate Professor, Higher School of Technology Essaouira, Cadi Ayyad University

**Morocco**

### Aakash Guglani

Senior Manager, Digital India Foundation

**India**

### Arina Gvozdyreva

Expert of the Human-Centered AI Center, Sberbank PJSC

**Russia**

### Rustam Hamdamov

Doctor of Technical Sciences, Professor, Head of the «Smart Systems and Internet of Things» Laboratory at the Research Institute for Digital Technology and AI Development under the Ministry of Digital Technologies of the Republic of Uzbekistan

**Uzbekistan**

### Dr. Yanio Hernández Heredia

PhD, Full Professor at the University of Informatics Sciences (UCI). Head of the Artificial Intelligence Research Group at UCI; President of KAINOS, S.A.

**Cuba**

### MSc. Denys Buedo Hidalgo

Director of Computerization at the Ministry of Higher Education

**Cuba**

### Prof. Dr. Essam Halim Houssein

Professor of Artificial Intelligence, Dean of Faculty of Computers & Information, Minia University

**Egypt**

### Evgeny Ilyushin

Assistant Professor at the Department of Information Security, Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, CAIO at Okko

**Russia**

### Dr. Leonel Iriarte

PhD, Senior researcher and Chief of the Artificial Intelligence Research Group at company DATYS

**Cuba**

### Dr. Mohanad Ali Mohammed Jawad

General Company for Air Navigation Services

**Iraq**

### Alexey Karpov

Head of the Speech and Multimodal Interfaces Laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS)

**Russia**

### Dr. Pritee Khanna

PhD, Professor, Computer Science and Engineering, and Dean of Research, Sponsored, and Consultancy Projects at PDPM Indian Institute of Information Technology, Design and Manufacturing

**India**

### DSc., Prof. Sergey Kolyubin

Professor, Head of BE2R Research Lab, Head of Robotics and AI MSc Program, Vice-Director for School of Computer Technologies and Control, ITMO

**Russia**

### Anton Konushin

PhD, Head of Spatial AI research group, AIRI Institute

**Russia**

### Ketan Kotecha

Dean of engineering symbiosis international university

**India**

### Alexander Krainov

Director of AI Technology Development at Yandex

**Russia**

### Maria Lapina

Associate Professor of the Department of Computational Mathematics and Cybernetics, Faculty of Mathematics and Computer Science North Caucasus Federal University

**Russia**

### Igor Lavrov

MD, PhD, Head of spinal cord neuromodulation group, Mayo Clinic, Rochester, US

**Russia**

### Prof. Xin Lin

Professor at East China Normal University

**China**

### Dr. Tran Quoc Long

PhD in Computer Science
Director of Institute for AI, VNU University of Engineering and Technology

**Vietnam**

### Narzillo Mamatov

Doctor of Technical Sciences, Professor, Head of the Department of Digital Technologies and AI, National Research University Tashkent Institute of Irrigation and Agricultural Mechanization Engineers

**Uzbekistan**

### Prof. Dr. Madina Mansurova

Head of the Department of AI & Big Data, al-Farabi Kazakh National University

**Kazakhstan**

### Sergey Markov

Director for AI Development, Sberbank

**Russia**

### Massimo Mecella

PhD, Professor, SAPIENZA Università di Roma, Dipartimento di Ingegneria Informatica, Automatica e Gestionale ANTONIO RUBERTI (DIAG)

**Italy**

### Vladimir Milovanović

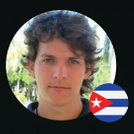
Professor of Electrical Engineering and Computer Science, University of Kragujevac

**Serbia**

### Dr. Gonzalo Ferrer Minguez

Associate Professor; Head, Mobile Robotics Lab, Artificial Intelligence Center, Skoltech

**Spain / Russia**



### Dr. Alejandro Piad Morffis

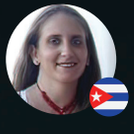
PhD, Computer Scientist, AI Researcher at the Faculty of Mathematics and Computer Science at the University of Havana

**Cuba**

### Khamidov Munis

Head of AI research laboratory, Associate professor in AI department of Samarkand State University

**Uzbekistan**



### Dr. Yailé Caballero Mota

PhD, Full Professor and Head of the Artificial Intelligence Research Group at the University of Camagüey. Member of the International Academy of Sciences

**Cuba**

### Prof. Dr. Alexey Naumov

Professor, Director of HSE AI and Digital Science Institute, HSE University

**Russia**

### Fayzulla Nazarov

PhD, Dean of the Faculty of Artificial Intelligence and Digital Technologies, Sharof Rashidov Samarkand State University named after Sharof Rashidov

**Uzbekistan**

### Sergey Nikolenko

Senior Researcher, PDMI RAS Head of the AI360 Educational Program «Mathematics of Machine Learning», Department of Mathematics and Computer Science, SPSU

**Russia**

### Prof. Aparajita Ojha

Professor of Computer Science and Engineering, Chief Investigator, Electronics and ICT Academy, PDPM Indian Institute of Information Technology, Design and Manufacturing

**India**

### Dr. Aleksandr Panov

Doctor of Physical and Mathematical Sciences, Head of Cognitive AI Systems Lab, AIRI Institute; Director of Center for Cognitive Modeling, MIPT

**Russia**

### Samprit Patel

Department of Computer Science and Engineering, Walchand Institute of Technology

**India**

### Pedro Yobanis Piñero Pérez

Subdirector of Centre for Research, Development and Innovation in Artificial Intelligence (CIDIIA) of the Central University of the East

**Dominican Republic**

### Dr. Rafael Bello Pérez

Full Professor. Director of the Computer Science Research Center. Head of the Artificial Intelligence Research Group at Central University of Las Villas (UCLV)

**Cuba**

### Dr. Raydel Montesino Perurena

PhD, Full Professor. Rector of the University of Informatics Sciences

**Cuba**

### Eduard Poghosyan

Professor, PhD and Doctoral in Mathematical Cybernetics, head of Cognitive Algorithms and Models Direction, Institute for Informatics and Automation problems of the National Academy of Sciences of the Republic of Armenia

**Armenia**

### Mohammed Abdul Qadeer

Associate Professor, Department of Computer Engineering, Aligarh Muslim University

**India**

### Dr. Feiwei Qin

Professor, School of Computer Science and Technology, Hangzhou Dianzi University

**China**

### Dr. Ernesto Estevez Rams

PhD, Senior Researcher, Full Professor, Faculty of Physics, University of Havana. Meritorious Academician of the Cuban Academy of Sciences

**Cuba**

### Samuel Rahimeto

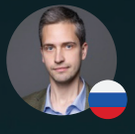
MSc, Director of Interpretable and NLP research Division, EAII

**Ethiopia**

### Akbar Rashidov

PhD, Associate Professor, Department of Artificial Intelligence and Information Systems, Samarkand State University named after Sharof Rashidov

**Uzbekistan**



### Oleg Rogov

PhD, Head of Reliable and Secure Intelligent Systems research group, AIRI Institute; Head of SAIL Lab, AIRI-MTUCI

**Russia**

### Dr. Samir Rustamov

PhD, Associate professor, ADA University

**Azerbaijan**

### Dr. Sri Safitri

Group Corporate Transformation Leader - Telkom Indonesia

**Indonesia**

### Dr. Mohammad Sajid

Assistant Professor Department of Computer Science

**India**

### Dr. Sergey Samsonov

Head of the International Laboratory of Stochastic Algorithms and High-Dimensional Inference, HSE University

**Russia**

### Andrey Savchenko

Prof., Doctor of Technical Sciences, PhD, Scientific director, Sber AI Lab

**Russia**

### Dr. Fazilov Shavkat

Head of the "AI and ML" Laboratory at the Institute of Digital Technologies and Artificial Intelligence, Doctor of Technical Sciences, Professor

**Uzbekistan**

### Alexei Shpilman

Managing Director - Head of Center "AI for Science", Sberbank PJSC

**Russia**

### Alexey Skrynnik

PhD, Head of RL agents research group at Cognitive AI systems Lab, AIRI Institute

**Russia**

### Dr. Neelakandan Subramani

Professor – Research, Department of Computer Science and Engineering, R.M.K Engineering College, Chennai - India

**India**

### Dodi Sudiana

Professor, Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia

**Indonesia**

### Ibragim Suleimenov

Almaty University of Power Engineering and Telecommunications

**Kazakhastan**

### Levon Aslanyan

Doctor of Physical and mathematical sciences, Head of Discrete Mathematics Department, Institute for Informatics and Automation Problems of the National Academy of Sciences of the Republic of Armenia

**Armenia**

### Alain Garofalo Hernandez

PhD, Chief Product Officer Avangenio SRL

**Cuba**

### Prof. Mbuyu Sumbwanyambe

Head of department University of South Africa

**Republic of South Africa**

### Dr. Elena Tutubalina

Doctor of Computer Sciences, Head of the Domain-Specific NLP research group, AIRI Institute; senior research scientist, ISP RAS

**Russia**

### Denis Turdakov

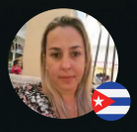
Head of the Research Center for Trusted AI, Ivannikov Institute for System Programming of the Russian Academy of Sciences

**Russia**

### Lev Utkin

DSc, Professor at the Higher School of Artificial Intelligence Technologies in Peter the Great St.Petersburg Polytechnic University

**Russia**



### Dra. Heydi Mendez Vazquez

PhD, Director and Senior Researcher at the Center for Advanced Technology Applications, Pattern Recognition and Data Mining (CENATAV)

**Cuba**

### Dr. Suilan Estevez Velarde

PhD, Dean of the Faculty of Mathematics and Computer Science at the University of Havana

**Cuba**

### Serestina Viriri

PhD, Professor (Computer Science). Head of Computer Vision and Machine Learning research group, UKZN, South Africa

**Republic of South Africa**

### Yelizaveta Vitulyova

Almaty University of Power Engineering and Telecommunications / Al-Farabi Kazakh National University

**Kazakhastan**

### Prof. Dianhui Wang

Director, Research Centre of AI for Engineering, Qingdao University of Science and Technology

**China**

### Dr. Riza Hammam

President of KORIKA (Artificial Intelligence Industry Research and Innovation Collaboration)

**Indonesia**

We express our sincere gratitude to Alexander Nikolaevich Gorban for his participation in this project. Alexander Nikolaevich was a great scientist whose work made significant contributions to the development of modern science.
His research spanned a wide range of fields — from statistical physics and non-equilibrium thermodynamics to machine learning and mathematical biology.

### Alexander Nikolaevich Gorban

Head of the AI, Data Analysis and Modeling Laboratory at Central University and Artificial Intelligence Research Institute

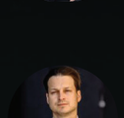
**Russia**

Eternal bright memory.

# PROJECT TEAM

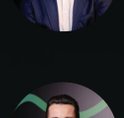
**Dr. Andrei Neznamov**

Managing Director of the Human-Centered AI Center, Sberbank PJSC, Secretary-General, AI Alliance Network



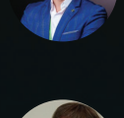
**Oleg Artyugin**

Executive Director of the AI Initiatives Implementation and Promotion Directorate, Sberbank PJSC



**Rustam Borovik**

Section head of AI Transformation of the Human-Centered AI Center, Sberbank PJSC



**Arseniy Lisov**

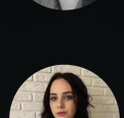
Section head of the Direction of the Strategic Agency for Support and Formation of AI Development

**Arina Gvozdyreva**

Expert of the Human-Centered AI Center, Sberbank PJSC

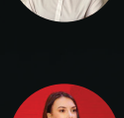
**Elvira Chache**

Executive Director of the Human-Centered AI Center, Sberbank PJSC



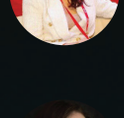
**Daria Churilova**

Section Head of the Human-Centered AI Center, Sberbank PJSC



**Sofia Fokina**

Section Head of the Human-Centered AI Center, Sberbank PJSC



**Yulia Zemtsova**

Leading Data Researcher of the Human-Centered AI Center, Sberbank PJSC

# INFORMATION ABOUT THE ORGANIZER — AI ALLIANCE NETWORK

AI ALLIANCE NETWORK

aianet.org

## AI Alliance Network

international community uniting associations in the field of AI

**20+** countries

**25+** associations in the field of AI

**7000+** affiliated partners